New Orleans JAZZED about Engineering Education

Paper ID #16199

A Laboratory Study of Student Usage of Worked-example Videos to Support Problem Solving

Dr. Edward J. Berger, Purdue University, West Lafayette

Edward Berger is an Associate Professor of Engineering Education and Mechanical Engineering at Purdue University, joining Purdue in August 2014. He has been teaching mechanics for nearly 20 years, and has worked extensively on the integration and assessment of specific technology interventions in mechanics classes. He was one of the co-leaders in 2013-2014 of the ASEE Virtual Community of Practice (VCP) for mechanics educators across the country.

Prof. Michael Wilson, Purdue University, West Lafayette

M.D. WILSON is a lecturer for the Krannert School of Management, the entrepreneur-in-residence for the Office of Future Engineers, and a Ph.D. candidate at Purdue University in the College of Engineering; his ”Pracademic” background combines rigorous research with practical experiences. Wilson started, sold, and consulted Fortune companies in the University-Industry entrepreneurial space for over twenty successful years. He earned a Bachelors of Science from the University of Massachusetts and a Masters from the University of Chicago; his broad research interests include Engineering Education, Network Science, and Modeling Human Sociometrics. Professor Wilson may be reached at wilsonmd@purdue.edu

©American Society for Engineering Education, 2016

A laboratory study of student usage of worked-example videos to support problem solving

Abstract

Despite the commonplace usage of video resources for engineering instruction, an understanding of precisely how students use such videos to support their problem solving and learning is incomplete. Researchers generally find that both students and faculty like using instructional videos (if 'well constructed'), especially in the format of so-called 'worked examples' in which an expert records a problem solution for learner consumption. Cognitive load theory (CLT) has successfully affirmed instructional worked-example interventions as more effective and efficient than problem solving in novice-phase skill acquisition. However, most worked-example studies look at pre/post performance on problem solving in which the worked-example is the intervention, rather than studying student use of the worked-example itself in great detail. This study begins to address the gap in our understanding of how students use worked-example videos to support their problem solving. In this laboratory-based research, we studied problem-solving processes of a group of 24 students enrolled in a required mechanical engineering sophomore-level course. In the experiment, students were presented a dynamics problem to be solved, provisioned with an equation sheet, an online calculator, and a video described as 'potentially useful.' Real-time data about student problem solving process and use of the video was captured via a Livescribe smartpen and a Mirametrix eye-gaze capture system (which captured their interactions with the video). Pre- and post-surveys about student attitudes about technology, perceptions of task difficulty, and academic transcript information are also included in the data set. Experimental videos and transcripts were coded for themes, and data about both task efficiency and task performance were extracted from the experimental evidence. Taken together, the results suggest that student usage of video resources can be broadly described by several archetypes, although in this study successful problem solution was possible regardless of archetype. These results will continue to inform academic coaching of students in our classes about optimal use of video resources.

Introduction

Assessments in sophomore-level mechanical engineering courses such as statics, dynamics, and thermodynamics, often emphasize problem solving, and indeed instruction is usually oriented around problem solving approaches and examples. In the last 10 years, instructional supports in the form of worked-example videos have become quite common, for two reasons. First, authoring tools for video creation continue to increase in power and ease of use, while simultaneously dropping in price. Second, the research on the worked-example effect[1]–[3] continues to support the notion that video-based worked examples can be effective instructional supports for novice learners. The coalescence of these two factors has led to the ubiquity of instructional videos available online across a huge range of topics.

The research on the effectiveness of worked examples is persuasive. The worked-example effect is a learning effect predicted by cognitive load theory (CLT)[4], [5]. Worked-examples are among the strongly-guided instructional strategies that reduce cognitive load in novices who learn by

observing experts solving problems. When used as part of instruction, worked-examples, compared to many other techniques, improve learning during skill acquisition[6]–[8]. The cognitive loads within a learners' working memory are induced by tasks, performance, and the mental effort invested[9]. Using well-structured multimedia-oriented instructional designs can reduce learners' extraneous cognitive loads[10]. Furthermore, learners use separate processing systems to process either visual (pictures) or auditory (verbal) representations of information. Mayer[11] developed a cognitive theory for multimedia learning, and it emphasizes clean design, complementary aural and visual information, careful attention to cognitive load of the learner. The model also demonstrates the influence that learning motivation and cognitive load, during the learning process, have on performance[12].

Right now, there is no widely agreed-upon approach to measuring cognitive load in an experimental environment. The four common methods presented in the literature all have their affordances and drawbacks. There are a variety of indirect measures related to, for instance, task performance[13], although they often suffer from confounding factors such as the experience of the learner and are therefore challenging to interpret. There are also secondary task designs in which response time to an external stimulus[4] (say, a request to click an on-screen button during completion of a load-inducing task), but their experimental design can be complication. There is a huge range of physiological indicators as well, including heart rate[14] and neural activity/EEG[15], but these approaches suffer from complicated experimental designs and (in some cases) prohibitive cost. One other approach, the one we use in this research, is a post-test subjective rating scale[16]. In particular, we use a modified version of the NASA-TLX task load index[17] that is easy to use, easy for test subjects to complete, and requires very little time. When balanced against the already complex experimental design used here (as described later), this brief and convenient measurement of workload was the best choice for our work. The NASA-TLX was initially developed for a broad range of tasks, including physically-intensive tasks. We have removed TLX items related to physical exertion, but have otherwise used the entire instrument as it was originally developed.

While the effectiveness of worked examples has been established, we currently do not fully understand exactly how students integrate worked examples into their study practices. Prior studies have largely viewed the use of worked examples as an intervention to support learning, with the metrics of the study related to pre-/post- gains in understanding or ability. We are interested in the details of how students use worked examples to solve problems, and there exists a gap in our current understanding of this facet of worked-example instruction. This gap in the literature inspires the broader research we are conducting, as well as the specific research questions considered in this paper:

- ***RQ1***: what are the necessary components of a laboratory experiment designed to probe student usage of worked examples in support of problem solving? *Working hypothesis:* we expect that real-time, video-based data—supplemented with pre- and post-surveys—will yield the most persuasive evidence about worked example use.
- ***RQ2***: to what extent do key metrics derived from the experiment predict academic performance on the example problem, or in the corresponding class? *Working hypothesis:* student usage of worked examples falls into several archetypes (i.e., usage patterns), but success in the experiment or in the corresponding class is possible regardless of worked-example usage archetype.

This paper describes findings from our first set of experiments designed to answer these two research questions.

Methods

Participants and recruitment

We recruited 24 subjects from a core mechanical engineering sophomore course, Dynamics, to participate in the study (12 from Spring 2015, 12 from Fall 2015). Subjects were recruited via in-person announcement and email, and no academic or demographic selection criteria were applied to the subject pool at the recruitment stage (i.e., no one was selected or disqualified based upon GPA, gender, or any other characteristic; to use the eye gaze equipment, there were several visual ability disqualifiers as detailed below). The subjects included three women and twenty-one men, with eight non-Mechanical Engineering majors. Our sample had an average GPA of 3.15 (the average sophomore student in ME has a GPA of about 3.39). Participants consented to engage in one 45-minute laboratory experiment, to complete pre- and post-surveys, and to allow the research team access to their academic transcript and admission data (SAT score, high school GPA, etc.). Subjects were compensated with a $20 Amazon gift card for their time. Participants enrolled in the dynamics class have access, as part of the class, to a large library of instructor-authored worked-example videos, so all participants had prior experience using such videos as part of their dynamics course. Although participants self-selected to participate in this experiment, their final course grades in the dynamics course roughly mirror the grade distribution for the class as a whole, as shown in Table 1.

Table 1. Comparison of participant performance and whole-class performance in dynamics.

Letter grade	Participants in this study	Spring 2015 dynamics	Fall 2015 dynamics
A	20.8%	18.6%	17.8%
B	29.2%	42.7%	34.8%
C	41.7%	28.1%	29.4%
D	4.2%	7.5%	8.9%
F	4.2%	3.0%	5.3%

Note: Columns do not add up to 100% due to rounding.

Laboratory experiment

After due consideration of our RQ1, we concluded that the experiment design required students to solve an actual dynamics problem under realistic (i.e., time-constrained) circumstances, and that the problem had to be non-trivial. We also decided to make three simultaneous, real-time measurements of student actions:

- Problem solving actions: student wrote their solution to the dynamics problem in a Livescribe notebook. We used an Echo smartpen system, which employs a small camera in the tip of the pen (along with specially printed paper) to record everything the student writes in a time-stamped way.
- Thought process: we asked students to follow a think-aloud protocol and describe their thoughts and actions verbally during the experiment. These verbal expressions were

audio recorded using the Echo smartpen, which automatically synchronizes the audio recording with the written work of the student.

- Worked example usage: students had access to a video-based worked example on a computer workstation directly in front of them. That workstation was equipped with Mirametrix eye gaze capture technology so that we captured when and where they were looking at the worked-example video. This action was captured in real-time via a screen recording on the workstation, and this video was synchronized with the Livescribe data.

To our knowledge, this is the first time such simultaneous measurements of student actions during problem solving have been made. These three actions capture the full range of observable student problem solving actions in one real-time experiment.

Participants completed a pre-survey of 7 items allowing them to self-report their level of comfort with and usage of the dynamics worked examples available as part of the dynamics course. Immediately after the problem-solving part of the experiment, but before taking the post-survey, students were asked 4 interview questions focusing on their experience of solving the problem with the support of the worked example. Participants then completed a 9-item post-survey allowing them to self-report perceived difficulty and level of effort required to solve the problem—a modified 5-item NASA TLX scale[17] that omitted the 'physical demand' dimension of task performance (7-point Likert scale), two items about perceived mental effort and problem difficulty (5-point Likert scale), one item about their perception of the usefulness of the worked-example video provided as part of the experiment, and one item to collect feedback about their perceptions of the video (too long, too short, too detailed, not detailed enough, etc.). The TLX results were calculated with a fixed weight scheme across the five dimensions, so the total TLX score for each participant was the average of their responses across the 5 dimensions on the 7-point scale. The five TLX questions were:

- How successful were you in solving this problem?
- How stressed, frustrated, annoyed, or discouraged were you while solving this problem?
- How hard did you have to work to solve this problem?
- You were allotted 25 minutes to solve this problem. Was this enough time to complete your solution?
- How mentally demanding was this problem?

For each question, a higher value of response on the 7-point scale indicates a more mentally demanding or stressful experience.

Experimental protocol

A total of 45 minutes was allocated to conduct the experiment from initial pre-screening through the signing of compensation forms. We asked subjects to solve a single, non-trivial dynamics problem that was congruent with their current progress in the course. In this case, the multi-part problem focused largely on proper usage of work-energy formulations to analyze particle kinetics. In both fall and spring semesters, participants completed the experiment within three weeks after covering the material in the dynamics class. A worked-example video, a single sheet of formulas, a Livescribe pen and booklet, and an onscreen calculator were provided.

Prior to the experiment, each subject was asked to declare their dominant writing hand (a setting for the Livescribe smartpen), whether they had bi -or trifocals or any of the following eye

conditions: glaucoma, cataracts, eye implants, or permanently dilated pupils. Best practices for using eye gaze technologies recommend discounting subjects with any of these conditions, as it may potentially corrupt the measurements. None of the subjects were disqualified during the pre-screen questions and all subjects were 18 years or older.

Upon signing the IRB consent form, the subjects completed the online pre-survey, which took less than three minutes to complete by each subject. Next, the subjects were asked to sit facing the 27-inch monitor and the eye-tracking device was positioned using a nine-point calibration test (generally 2-3 minutes for calibration). Overhead lighting was constant and each experiment occurred between 9 a.m. and 12-noon facing west. The subjects' backs faced the laboratory door to minimize distractions from the adjoining hallway.

Participants were given 25 minutes to solve a single (multi-part) dynamics problem, and the video available to them on the computer workstation was described as 'potentially helpful'. Participants then completed the problem to the extent they could in the 25-minute time limit, and used the video in whatever way they wanted to during the experiment (including not using the video at all). They were then asked 4 questions by the researcher, followed by the 9-item online post-survey. Participants were then thanked and compensated for their time, and the experiment ended. In all cases, the total time for this experiment was less than 45 minutes.

We used two different problems in this research with identical underlying mathematics, but the problem contexts were different. Because the experiment was conducted across two semesters, and because we provided participants a graded copy of their work (almost all participants asked for their grade), we were somewhat concerned about potential contamination of the applicant pool in the Fall 2015 semester. So we developed a new problem context that required identical equations, thought processes, and procedures to solve. To the participants, who are novices, the two problems looked entirely different. However, experts understand that the problems were actually identical. This difference in problem between the two semesters over which data was collected is a potential confounding factor when comparing absolute performance on the experimental task. However, in extracting information about how students solve problems, and how they integrate worked-example videos into their problem solving process, we expect this difference in problem to have a negligible effect.

Data Analysis

Quantitative data were collected from multiple sources, including the pre- and post-surveys, academic transcripts, admission data, and problem performance. Each problem solution was graded according to a rubric and based entirely upon the written contents of the solution. As such, the experimental problem was graded precisely as a written homework submission or exam would be graded. The problem contained three parts, and the grade for each part was recorded as part of the master data set. Various metrics extracted from the experimental videos themselves, as described below, were also included as quantitative data in the master data set. Quantitative data were analyzed using and R[18].

Qualitative data analysis proceeded as follows. The videos (Livescribe and eye gaze) were temporally synchronized with the experimental audio from the Livescribe pen into a composite video showing a complete record of student actions during the experiment. A screenshot of such

a composite video is shown in Figure 1. For each composite video, a special transcript was constructed, and this transcript captured four categories of information:

- Verbal information: participant think-aloud transcription
- Problem solution events: key steps in the solution, and their duration (example: drawing a free body diagram, writing a specific equation, or performing algebraic operations)
- Video events: playing the video, searching through the video, or (most importantly) watching the video (as detected using the eye gaze data)
- Affective events: participants audibly expressing positive affect ("I think this is right"), negative affect (a deep, frustrated sigh, or something like "I'm confused and not sure what to do here…"), or neutral affect (defined as an audible affect that is neither clearly positive nor clearly negative)

The experiment transcript is therefore not simply a transcription of the participant's verbal expressions during the experiment, but instead it is a composite record of verbalized expressions and problem solving actions in the same document.

The transcript and video were then imported into nVivo and coded for themes according to a 'node' structure developed prior to data analysis. nVivo is a qualitative data analysis software package produced by QSR International. It has been designed for qualitative researchers working with very rich text-based and/or multimedia information, where deep levels of analysis on small or large volumes of data are required. In nVivo, a 'node' is simply a specific category into which an observation fits, and we defined a multi-level node structure that has three high-level nodes (affect events, solution events, and video events), with many sub-nodes that provide granularity to the coding.

The node structure is shown in Table 2, and it enables us to capture the four categories of information described above. It is important to note that since this node structure was developed and used to analyze the data presented here, we have convened an expert panel of 4 veteran dynamics instructors to critique the node structure. Based upon their feedback, we are currently preparing a slightly revised node structure for use in all future data analysis. It is also important to recognize that the granularity in any node structure seeks to balance the ability to capture the salient features of the solution without containing so many nodes as to be overwhelming. We consciously made the choice to focus on process-oriented features of problem solving (i.e., specific fundamental actions students took) so that our node structure was versatile enough to allow analysis of solutions to many different kinds of problems and could be used for continuing research.

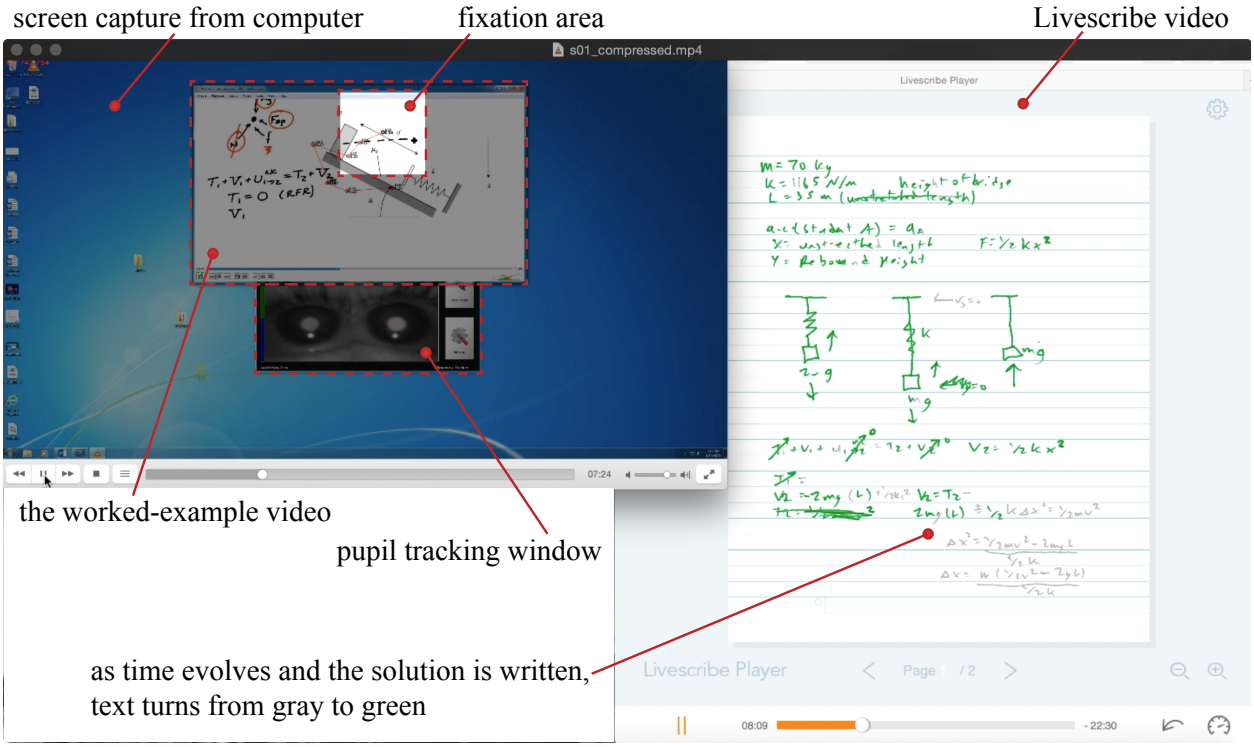

Figure 1. Composite video of one experiment showing the Livescribe video temporally synchronized with the eye gaze video. The figure shows the worked-example video window, the pupil tracking window, and the fixation location.

Table 2. Node structure for analysis of composite experimental video and transcript.

Event	Description
Affect Events	
negative	participant expresses doubt, frustration, or confusion
positive	participant expresses confidence or optimism
neutral	audible event that is identifiably neither positive nor negative
Solution Events	
breakthrough	participant pauses, seems stuck, needs to work hard, and then eventually unlocks how to move the solution forward
calculation	participant is actively using the calculator
decision	participant makes a decision; examples: choosing a coordinate system and orientation, a datum for energy calculations, or an analysis method
draw FBD	participant draws a free body diagram
forward reasoning	participant is actively moving the solution forward; includes: thinking through next steps, performing algebraic manipulations to isolate quantities of interest, drawing diagrams
metacognitive	participant is engaging in a self-check of their progress by questioning approach, process, or result
mis-step	participant makes an error in their solution
mis-step correction	participant corrects a previous mis-step
read problem statement	participant is reading the problem to understand the nature of

	the problem, the givens and unknowns, etc.
solution, part (a)	participant is solving part (a) of the problem
solution, part (b)	participant is solving part (b) of the problem
solution, part (c)	participant is solving part (c) of the problem
write F = ma	participant writes and works through the kinetic equation
write work-energy	participant writes and works through the W-E equation
task set-up	participant is defining variables, writing down given information, and otherwise preparing to begin the solution
unnecessary step	participant makes an unnecessary step in the solution and writes an equation or solves for a quantity that is not needed for the solution; unnecessary steps are not always mis-steps
Video Events	
video play	video is playing
video watch (fixation)	participant is actively watching the video
video search	participant is searching through the video
video re-watch	participant is watching a part of the video that he/she has already watched at least once before

Using nVivo to analyze the data allows us to perform the crucial step of extracting quantitative information from the experimental composite video, and this quantitative data can be used in further statistical analysis. This is most clearly illustrated using the concept of 'coding stripes', which are time-stamped visualizations of the coding of experimental information onto the node structure. Figure 2 shows an example set of coding stripes from one of our experiments, and it clearly illustrates the time-based parameters that can now be extracted from the experimental composite videos. From this information, nVivo allows us, via its query tools, to easily report out time parameters such as: video play frequency and time (in Figure 2, 9 play events totaling 09:50 min.), fixation frequency and time (13 events, 06:09 min.), and frequency and time of forward reasoning events (28 events, 08:10 min.).We also see that this participant spent 19:37 min. on the solution to part (a) of the problem, which coincides with a large amount of reasoning, video playing and fixation, and finally a breakthrough that enables the final calculation for part (a). This participant received full credit for part (a) of the problem, but failed to complete either part (b) or (c). In particular, a mis-step in part (b) [which amounted to using the *energy* expression $1/2\ kx^2$ to represent spring *force* on the free body diagram] prevented this subject from completing part (b) correctly.

[An interesting aside for this participant is that there was an attempt to troubleshoot the solution to part (b) by examining the units present in the kinetic equation. The participant realized that the units were inconsistent (a mixture of force and energy), but was unable to fully understand and repair this mis-step in the time allotted.]

The nVivo-derived quantitative metrics were then inserted into the master data set and used for various statistical analyses. In the remainder of the paper, we review several of our findings so far, focusing on ways we can link pre-survey, post-survey, grade, and experimental data.

Results

We have developed a laboratory experiment to explore how students solve problems with the support of worked-example videos, but to evaluate whether the experiment meets our original goals (our RQ1), we need to consider the usefulness of the data we have collected. Therefore, we begin the results section by considering a variety of metrics related to RQ2, then follow with a discussion of the extent to which the experiment we have developed meets our original goals.

RQ2: to what extent do key metrics derived from the experiment predict academic performance on the example problem, or in the corresponding class?

Participants completed the experiment with aggregate mean score of 14.95 (standard deviation 4.67) out of a possible 20 points. Students typically lost points for mis-steps in their solution around application of the appropriate analyses (work-energy, or Newtonian kinetics) or for the kinds of careless errors that often arise in a time-constrained task (algebra or calculation errors). Score on the experimental problem shows essentially no correlation to course grade, with Spearman correlation coefficients not rising to statistical significance for the total population of participants. This is not unexpected, however. In the same way that a single piece of graded work in the context of a course does not successfully predict the student's final course grade, their performance on a single problem in our laboratory environment probably will not predict their final grade either. There are many reasons why a student might or might not perform well on our experimental problem on a particular day in our lab, so a weak correlation between performance on the experimental problem and final course grade is expected.

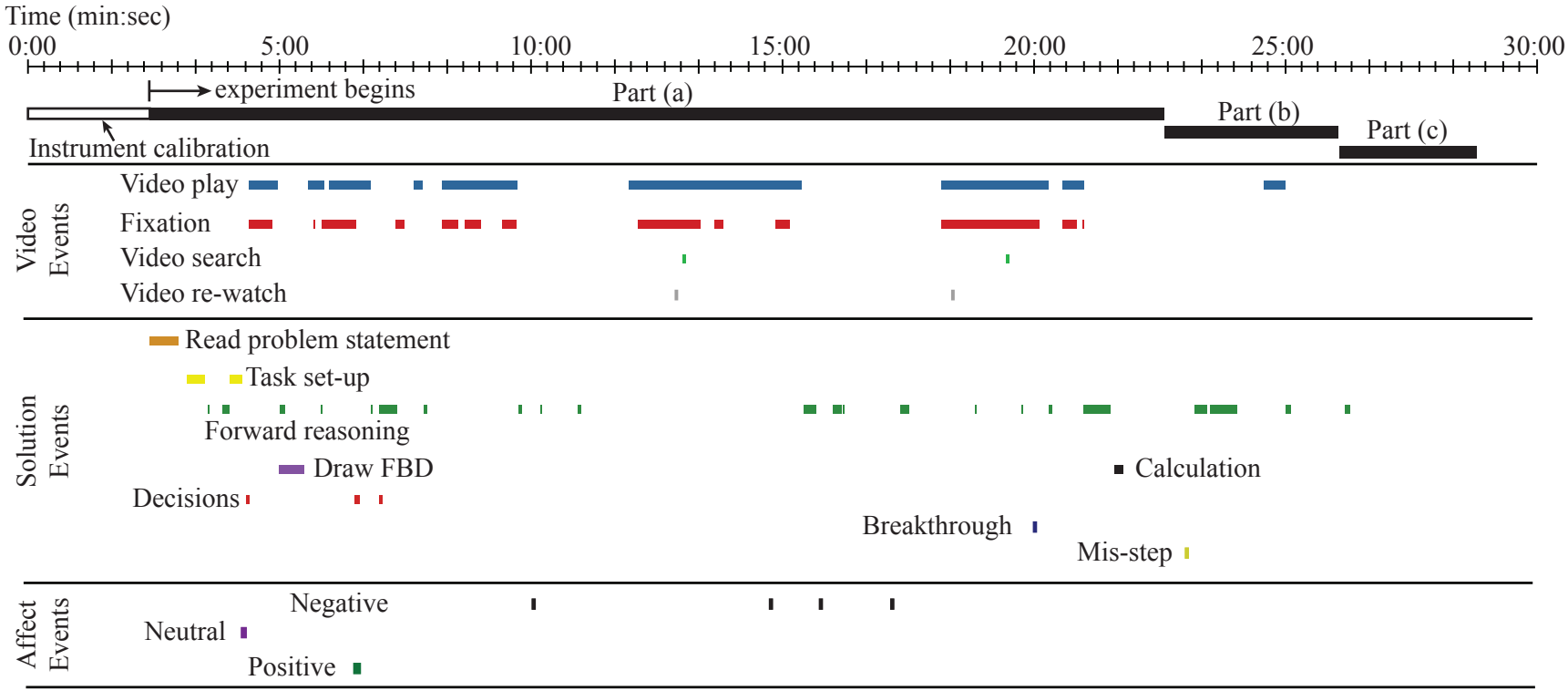

Figure 2. Coding stripes for experimental composite video showing time-stamped events. From this coding, quantitative time-based metrics are extracted from each experiment and used in subsequent statistical analysis.

So in addition to task performance, we also focused on several interior elements of the experimental data, especially patterns of video usage. Our initial hypothesis for RQ2 has, embedded in it, the idea that there are many possible ways for students to exhibit successful problem solving. We explored relationships among experimental time-based parameters (video play time, video watch time), experimental problem performance and overall grade in the dynamics course, and pre- and post-survey results.

Observation 1: Video fixation time and video play time were nearly equal

For all but one participant, video fixation time was nearly identical to video play time. The implication is that participants attempted to use the video available during the experiment in targeted ways, rather than having it play in the background as they worked the problem. This experimental observation is consistent with what participants reported on the pre-survey: 18 out of 25 reported that their typical use of the videos available as part of the dynamics course was a targeted attempt to understand a specific feature of the solution, rather than watching the video from beginning to end. During the experiment, we saw participants engage in a number of different video watch and search strategies. Some started their problem solving process by watching the video, perhaps in the hope or expectation that the problem solved in the video we provided to them would be very reflective of the problem they were asked to solve. Others worked the problem on paper and consulted the video only when unsure about a step, or sometimes to confirm that their approach was correct.

Observation 2: High-achieving students watched the video during the experiment less

Figure 3 shows fixation time and dynamics course grade as a function of performance on the problem completed during the laboratory experiment. There is a visible cluster of students who performed well in the course, performed well on the experimental problem, and had low fixation time. This observation is consistent with the notion that high-achieving students need fewer instructional supports than other students—this is why they are high achieving. Even for C-students who scored well on the experimental problem, their fixation time tended to be much higher than the A- and B-students, on the order of 1.5-2 times higher (with the exception of one student who did not use the video in the experiment at all, earned a perfect score on the experimental problem, but earned a C in the course). The observation that video usage peaks with students whose grades are in the C range has been suggested before[19], and our pre-survey results confirm this as well. On average, the C-students in this study report watching about twice as many videos during the semester as the A-students, and about 50% more than B-students. Our experimental population contained only one participant who earned a D in dynamics, and one who earned an F in dynamics; we therefore cannot say anything conclusive about their video-using behaviors.

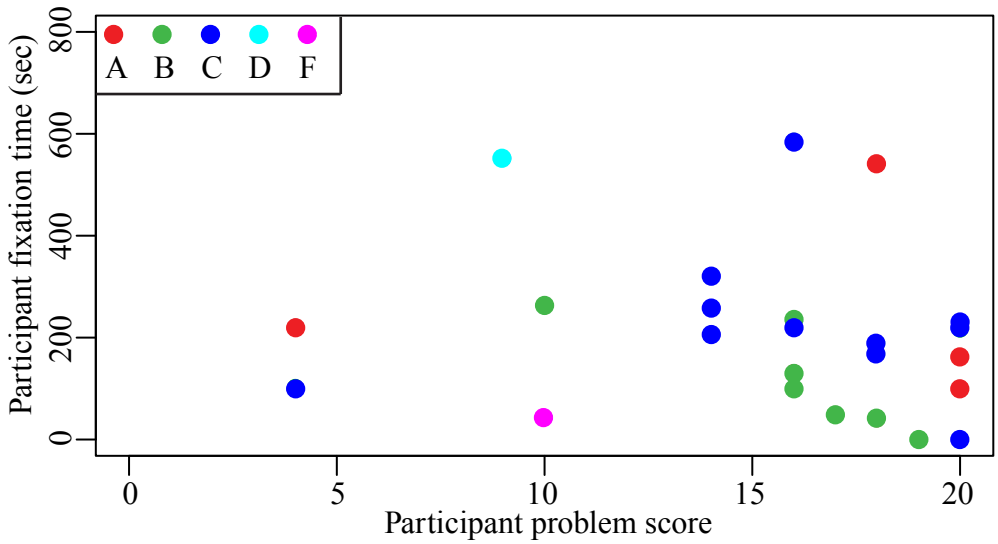

Figure 3. Fixation time in laboratory experiment and dynamics course grade as a function of score on the experimental problem.

Observation 3: Participants who perceived the problem to be challenging used the video more and performed worse

The post-survey TLX score was examined in light of experimental problem performance and fixation time, and in Figure 4 the data are colored by dynamics course grade. The regression line on the Figure 4(a) illustrates the inverse relationship between TLX score and problem score (meaning that students who perceived the task to be harder earned a worse score on the experimental problem; this is a statistically significant model of student performance, with $R^2 = 0.49$ and $p < 0.0001$). Figure 4(b) shows an increasing trend of fixation time with TLX score, meaning that students who perceived the experimental task to be harder generally used the video more (this regression model is not statistically significant).

In Figure 5 the experimental data points are colored by the student's perception of their experimental task success as reported on the post-survey (a 7-point Likert-scale item ranging from 'perfect' to 'a failure'); this question was one of the 5 elements of the TLX. Figure 5(a) neatly shows that student perceptions about their performance on the experimental task were generally correct and trended very closely with their TLX score. Students who perceived the task to be more difficult were generally less confident about their performance, and their grade on the experimental problem reflects this. The conclusion here is that students are reasonably good at assessing their own performance on the single problem they solved during the laboratory experiment. Figure 5(b) shows another view of the data that reinforces the notion that participants who perceived the task to be harder, and were less confident about their performance, used the video more during the experiment.

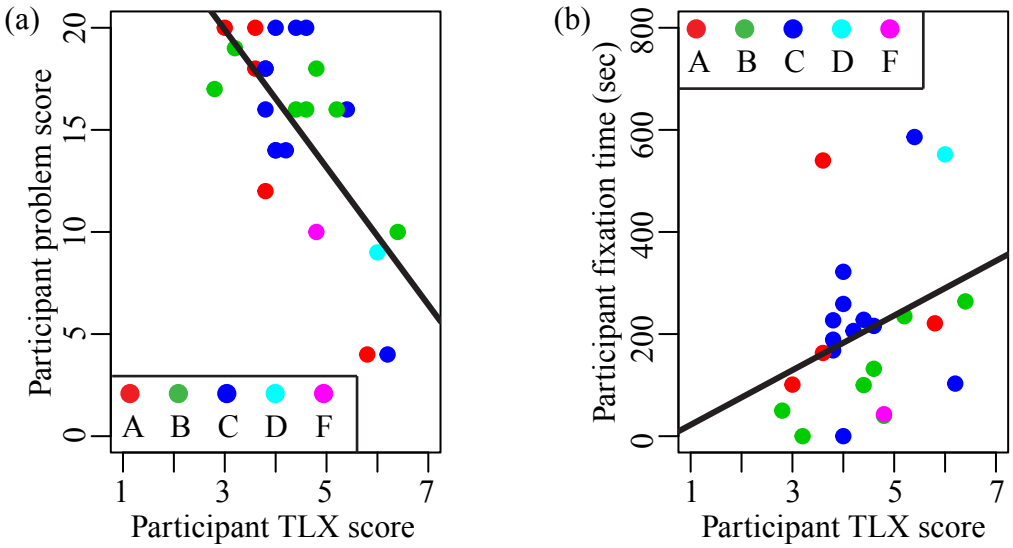

Figure 4. Experimental task performance (a) and fixation time (b) as a function of TLX score. Data points are colored according to each participant's final course grade in dynamics.

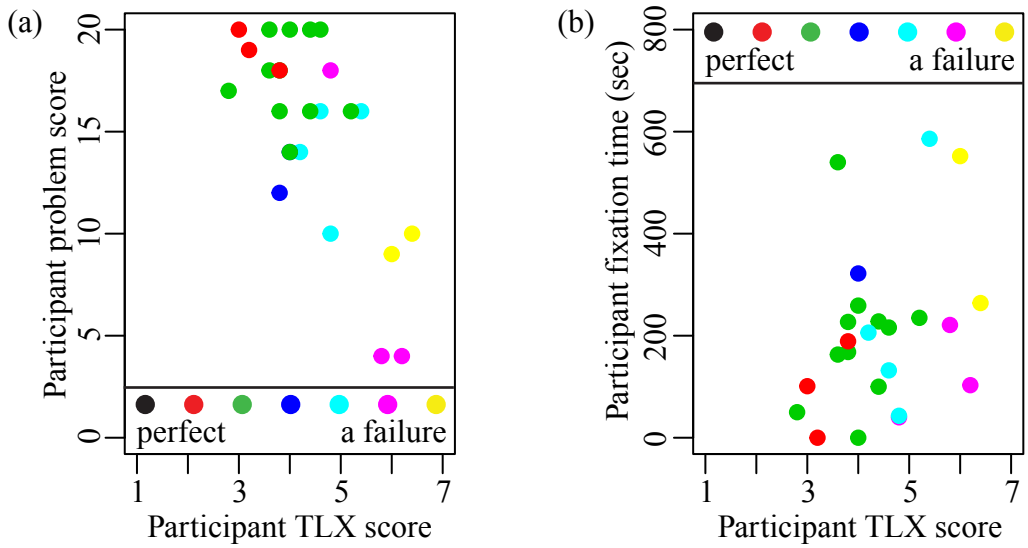

Figure 5. Experimental task performance (a) and fixation time (b) as a function of TLX score. Data points are colored according to each participant's perceived success in solving the experimental problem.

RQ1: what are the necessary components of a laboratory experiment designed to probe student usage of worked examples in support of problem solving?

Now that we have reviewed some of the results related to RQ2, we return to RQ1 and the question of a laboratory experiment to explore video usage. The results presented above give us confidence that the experiment we developed for this preliminary study contains many of the salient elements important for answering questions about how students solve problems with access to worked-example videos.

Although we have focused on a small subset of the available data collected during the experiment, our ability to resolve specific actions students take during problem solving is substantially advanced by the real-time, video-based measurements described here. The think-aloud protocol enables a window into the participant's mind, and also allows us to collect affect information from participants. The pre- and post-survey data also seem to be very important and complementary to the measurements made during the experiment. In particular, the TLX task workload rating gives us important feedback about the participant's perception of their mental effort and could lead toward useful characterizations of cognitive load during problem solving.

Nonetheless, we plan several improvements to the experiment to remedy specific issues encountered during this first round of data collection. The promise of eye gaze technology is that we can know, with very high resolution, exactly where the participant is looking on the computer screen. So we should be able to tell whether a participant is looking at the figure, the problem statement, or a particular equation while fixated on the worked-example video during the experiment. However, because the participant frequently moves their head back and forth—looking at the computer screen, then their written work, and back—we are not confident that the eye gaze system calibration or pupil tracking is robust against those kinds of discontinuous use patterns. As a result, in the current experiment, we can confidently determine when a participant watched the video on screen (by monitoring the pupil tracking window in Figure 1), but we cannot confidently say exactly what the participant was looking at (we do not have confidence in the location of the fixation window). We continue to examine potential remedies for this situation, but for the current set of results we simply cannot say with confidence exactly what the participant was looking at in the worked-example video. This prevents us from making any statements about the relative value of, say, clear and detailed hand-drawn sketches, very detailed equation derivations, or more general expository information. Developing an approach to gain further confidence in the specific location of the fixation window would substantially extend this research and be of general value to the community.

We also expect to use a stationary mounted camera on the desk near the participant to capture a third-person perspective on the experiment. This additional camera view will allow us to capture facial expressions and supplement the other video evidence with confirmatory information throughout the experiment.

Despite these limitations and the lessons learned through this preliminary data collection, we feel confident that the experimental system and protocol proposed here can be used—with some of the refinements described above—for detailed analysis of student problem solving in the presence of worked-example videos. Our on-going work includes not only experimental refinements, but also much more detailed analysis of the data we have already collected.

Conclusions

In this paper, we describe a novel experimental approach and protocol that uses real-time video and audio data collection to examine student problem solving using worked-example videos for support. The hardware, experimental protocol, and data analysis approach together define a powerful approach to surveillance of student problem solving behaviors, and the 24 subjects who participated in this first round of experiments have demonstrated the affordances of the approach, as well as some limitations. The time-stamped nature of the measurements allows us to extract a

huge range of time-based parameters (Figure 2) that characterize the process by which students solve problems, including their use of the worked-example video. The data analysis protocol alone is a significant step forward in our ability to understand problem-solving processes and quantify, in terms of both frequency and total time, how students stitch together a series of discrete choices and actions into their overall solution to a dynamics problem.

Our preliminary data analysis illustrates several important trends of video usage during problem solving and its relationship to both task performance and perceived task difficulty. Students who believe the problem is more difficult used the provided worked-example video more, were less confident in their performance, and actually performed worse. While there was not a strong relationship between performance on the experimental problem and overall dynamics course grade (and none was expected), these relationships on a per-problem basis begin to build evidence of successful problem-solving strategies for students. While we cannot, based upon the 24 participants in the current study and the data analysis we have completed so far, conclusively describe problem-solving archetypes, this experimental and data analysis protocol is a significant step forward toward that goal.

Acknowledgement
The authors gratefully acknowledge the financial support of Purdue University for equipment, laboratory space, and student research assistantship funds. We also appreciate the willing participation of the student subjects who completed the experiment and provided incredibly useful feedback about potential refinements to our methods.

References

[1] J. Sweller, "The worked example effect and human cognition," *Learn. Instr.*, vol. 16, no. 2, pp. 165–169, Apr. 2006.

[2] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, "The expertise reversal effect," *Educ. Psychol.*, vol. 38, no. 1, pp. 23–31, 2003.

[3] R. Moreno, M. Reisslein, and G. Ozogul, "Optimizing Worked-Example Instruction in Electrical Engineering: The Role of Fading and Feedback during Problem-Solving Practice," *J. Eng. Educ.*, vol. 98, no. 1, pp. 83–92, 2009.

[4] J. Sweller, "Cognitive load during problem solving: effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, 1988.

[5] J. Sweller and G. A. Cooper, "The use of worked examples as a substitute for problem solving in learning algebra," *Cogn. Instr.*, vol. 2, no. 1, pp. 59–89, 1985.

[6] T. van Gog, F. Paas, J. J. G. van Merriënboer, and P. Witte, "Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting.," *J. Exp. Psychol. Appl.*, vol. 11, no. 4, pp. 237–44, Dec. 2005.

[7] R. Moreno, M. Reisslein, and G. Delgoda, "Toward a fundamental understanding of worked example instruction: Impact of means-ends practice, backward/forward fading, and adaptivity," *Proceedings. Front. Educ. 36th Annu. Conf.*, pp. 5–10, 2006.

[8] A. Renkl, "Learning from Worked-Out Examples: A Study on Individual Differences," *Cogn. Sci.*, vol. 21, no. 1, pp. 1–29, Jan. 1997.

[9] F. Paas and T. van Gog, "Optimising worked example instruction: Different ways to increase germane cognitive load," *Learn. Instr.*, vol. 16, no. 2, pp. 87–91, Apr. 2006.

[10] R. E. Mayer and R. Moreno, "Nine ways to reduce cognitive load in multimedia learning," *Educ. Psychol.*, vol. 38, no. 1, pp. 43–52, 2003.

[11] R. E. Mayer, “Multimedia learning,” in *Psychology of Learning and Motivation*, vol. 41, B. H. Ross, Ed. Academic Press, 2002, pp. 85–139.
[12] J. M. Keller, “An Integrative Theory of Motivation, Volition, and Performance,” *Technol. Instr. Cogn. Learn.*, vol. 6, no. 2, pp. 79–104, 2008.
[13] P. Ayres, “Using subjective measures to detect variations of intrinsic cognitive load within problems,” *Learn. Instr.*, vol. 16, no. 5, pp. 389–400, 2006.
[14] F. G. W. C. Paas and J. J. G. Van Merriënboer, “Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach,” *J. Educ. Psychol.*, vol. 86, no. 1, pp. 122–133, 1994.
[15] P. Antonenko, F. Paas, R. Grabner, and T. van Gog, “Using Electroencephalography to Measure Cognitive Load,” *Educ. Psychol. Rev.*, vol. 22, no. 4, pp. 425–438, 2010.
[16] A. Schmeck, M. Opfermann, T. van Gog, F. Paas, and D. Leutner, “Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings,” *Instr. Sci.*, vol. 43, no. 1, pp. 93–114, Aug. 2014.
[17] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results and Empirical and Theoretical Research,” in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds. Amsterdam: North Holland Press, 1988.
[18] R Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria, 2012.
[19] E. J. Berger and E. Pan, “Video Resources and Peer Collaboration in Engineering Mechanics : Impact and Usage Across Learning Outcomes Video Resources and Peer Collaboration in Engineering,” in *Proceedings of the 122nd ASEE Annual Conference and Exposition*, 2015, p. Paper ID #12100.