J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 2, Number 2 · August 2003

An Exploratory Study to Examine the Feasibility of Measuring Problem-Solving Processes Using a Click-Through Interface

Gregory K. W. K. Chung and Eva L. Baker

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 2, Number 2

An Exploratory Study to Examine the Feasibility of Measuring Problem-Solving Processes Using a Click-Through Interface

Gregory K. W. K. Chung and Eva L. Baker

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Kathleen O'Connor
Design and Layout: Thomas Hoffmann

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment, 2*(2). Available from http://www.jtla.org

Abstract:

In this study we investigated the feasibility of a novel user interface to support the measurement of problem-solving processes. Our research questions addressed the use of a "click-through" interface to measure the "generate-and-test" problem-solving process for a design problem. A click-through interface requires the user to explicitly perform an online action (e.g., to view time, the user has to click on a "time" icon). This interface allowed us to measure participants' intentional acts. Freshman college students were given the task of modifying a given, computer-interactive bicycle pump to satisfy performance requirements. The simulation interface provided participants with point-and-click access to controls to modify pump parameters, to run the simulation, to view important information, and to attempt to solve the task. Lag sequential analyses of participants' problem-solving processes over time showed cyclical behavior consistent with the generate-and-test strategy of modifying the pump design, running the simulation, viewing the information, and then either modifying the design or attempting to solve the problem and then modifying the design again. This behavior set was remarkably stable, with most lag 1 associations greater than .80. Our approach to measuring problem-solving processes appears feasible and promising, but more work is needed to gather additional validity evidence.

J·T·L·A

An Exploratory Study to Examine the Feasibility of Measuring Problem-Solving Processes Using a Click-Through Interface

One of the most promising aspects of computer-based assessments is the capability to provide students with complex tasks and the flexibility to unobtrusively measure student learning and problem-solving processes in real-time (Baker & Mayer, 1999; Bennett, 1999; Chung & Baker, 2003; Clauser, 2000; Huff & Sireci, 2001; National Research Council, 2001). Assessing problem solving typically involves providing students with the content, asking them to use the information in a novel way, and scoring the quality of the response. The cognitive demands of the task can range from requiring students to solve a simple problem that has a single answer, to requiring students to design solutions to problems that are ill-defined and can have multiple solutions. Examples of computer-based problem-solving tasks include requiring participants to design simple digital circuits (e.g., Katz & James, 1998), to find relevant information on the Web to help improve their understanding of environmental science given feedback (Schacter, Herl, Chung, Dennis, & O'Neil, 1999), and to use data as evidence to successively eliminate candidate solutions (Stevens, Ikeda, Casillas, Palacio-Cayetano, & Clyman, 1999).

Most methods of assessing problem-solving skills are based almost exclusively on paper-based formats and focus on outcomes. Outcome performance is assumed to be an index of students' understanding and competence in problem solving. However, outcome measures are by definition incomplete measures of student learning and understanding because they do not provide direct evidence of the kinds of processes students use throughout a task. Even when process data are collected, for example by videotaping students while they attempt to solve a problem, the effort involved is extremely labor-intensive or the data may suffer from inaccuracies of self-reporting (e.g., querying students throughout the task; asking students to recall what they did during the task). As a result of these difficulties in collecting process data, evidence is rarely gathered on the processes that students use, during learning and during problem solving, that lead to the observed performance.

Administering problem-solving tasks online is an attractive alternative to using paper-based formats for the purpose of measuring problem-solving processes (Baker & Mayer, 1999; Baker & O'Neil, 2002; Chung, de Vries, Cheak, Stevens, & Bewley, 2002; O'Neil, 1999). The online context provides a testbed to observe users' actions and draw inferences about their problem-solving processes.

The value of assessing problem-solving processes is that data on such processes can provide evidence of what a learner is doing while carrying out a task. Process evidence can be used, for example, to help evaluate the extent to which a task evokes expected problem-solving behaviors (e.g., in theory testing), help explain performance differences between subgroups (e.g., high and low knowledge individuals; high and low performers), or aid in task validation (American Educational Research Association & American Psychological Association, 1999; National Research Council, 2001). Thus, measures of problem-solving processes, when used in conjunction with measures of knowledge and problem-solving performance, can provide a more comprehensive picture of the learner.

However, two key issues need to be addressed that bear directly on the validity of inferences drawn about students' problem-solving processes. The first issue is deciding what to measure and at what grain size. Online tasks can be viewed as a kind of software testbed, and as such, *software sensors* can be embedded in the online task to gather telemetry about the state of the user at multiple grain sizes (e.g., from measuring participants' individual mouse clicks to measuring the number of problems solved). The second issue is related to validity – how is the construct of interest measured via software sensors, and how do you know it reflects the outcome of substantive cognitive processing?

Current Study

One challenge faced in measuring cognitive processes in online environments is that cognitive processes cannot be directly observed. What can be measured are the outcomes of cognition (i.e., problem-solving performance measures) and the online activity of the participant as he or she carries out the task. In the current study, a special user interface was designed and tested, and the main objective of the design was to support the direct observation of problem-solving processes. We used a multimedia-based simulation authoring system to develop the problem-solving task. Student responses and student interactions with the simulation were continuously logged (Munro et al., 1997; Towne et al., 1990). The cause-effect system that was simulated was a bicycle pump, a topic used in previous studies (Herl et al., 1999; Mayer & Gallini, 1990).

Design Problems and the Generate-and-Test Problem-Solving Strategy

The approach we took in this study was to examine problem-solving processes in the context of a design problem. In a design problem, participants need to devise a solution to satisfy a set of constraints. There are usually a number of constraints that can vary; thus, there are usually a number of ways to design a solution (Atman, Chimka, Bursic, & Nachtmann, 1999; Katz & James, 1998; Mullins, Atman, & Shuman, 1999). Participants solving design problems have been observed, via think-aloud protocols, to commonly use a generate-and-test strategy (e.g., Katz & James, 1998; Nhouyvanisvong & Katz, 1998).

In a generate-and-test strategy, the problem solver generates a candidate solution to the problem, tests whether the solution satisfies the constraints of the problem, and if not, generates another solution. Figure 1 shows a flowchart of the model of the generate-and-test strategy, adopted from Nhouyvanisvong and Katz (1998) and updated for the current study. As displayed in Figure 1, the simulation software used in this study permits an additional decision step (i.e., the participant can attempt to explicitly solve the design problem).

Figure 1

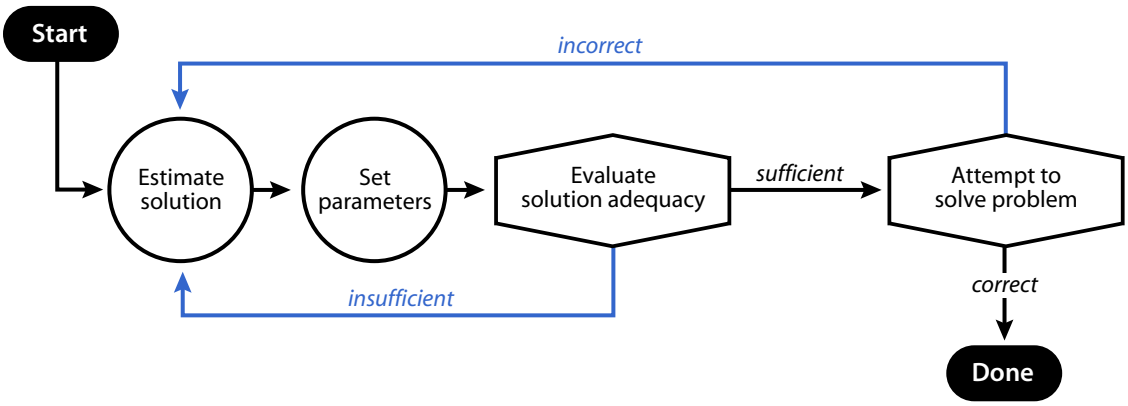

Figure 1. A flowchart of the generate-and-test strategy for the current study.

For example, Katz and James (1998) compared participants' problem-solving strategies used to solve two design problems to problem-solving strategies used to solve two analysis problems. Of relevance to the current study are the findings for participants solving design problems. In their analysis of participants' verbal protocols, Katz and James observed that participants solving design problems commonly used the generate-and-test strategy. In the context of the particular design problems (simple digital logic design problems), one problem-solving strategy used by participants was to first generate a potential solution, and then mentally "run" the circuit by supplying sample input signals and evaluating the output signal against the circuit criteria. If the output signal did not match the design criteria, the participant would redesign (or regenerate) the circuit and repeat the process.

The findings of Katz and James (1998) are important because they identify a set of processes that learners use to solve design problems. In this study we attempted to automate the measurement of these processes via measures derived from users' interaction with the user interface.

Click-Through Interface

To facilitate the direct measurement of participants' problem-solving strategies when attempting to solve a design problem, we developed a special "click-through" interface for participants. Participants were given access to all relevant information

about the design task, but to access the information they had to explicitly click on an interface element. For example, the task in this study involved modifying the features of a computer-interactive bicycle pump. The user interface made obvious to the participant where the information could be seen, but in order to view the information the user had to explicitly click on a button and hold the button down to view the current value.

Thus, participants had to click through the interface to see information they desired to see, imposing a cost on the user. Our assumption was participants would absorb the cost only if they intended to view the information and they judged the information to be relevant to helping them solve the problem. Our interface was designed for assessment purposes and intentionally sacrificed ease of use for utility of measurement.

Our research questions for this study were (a) to what extent does the click-through user interface capture participants' generate-and-test problem-solving processes for a design problem? and (b) to what extent does the click-through user interface interfere with participants' task performance?

Method

Participants

Participants were recruited from a major university and a community college in southern California. While 31 participants were recruited, 5 participants were dropped due to equipment failure and 1 was dropped due to task non-compliance. The participants were undergraduate freshmen and had participated in a previous study. Of the remaining 25 participants, there were 12 males and 12 females (1 participant did not report gender); and 9 Asian American, 6 Latino, 4 White, 3 biracial, and 1 African American participants (2 participants did not report ethnicity). The mean high school GPA was 3.86 ($SD = 0.40$, $n = 20$), and the mean SAT I math and verbal scores were 614 ($SD = 78$, $n = 16$) and 571 ($SD = 95$, $n = 16$), respectively. Participants were also administered a measure of scientific reasoning ($M = 12.84$, $SD = 4.54$, $N = 25$). In general, participants' self-reported familiarity with the content of the simulation was low. The majority of participants reported not being familiar with bicycle pumps and having little or no experience with mechanical devices. Participants were paid $30 for participating.

Task

Design Task

The simulation task required participants to solve three pump design problems. Participants were presented with a "working" computer-interactive pump and they were instructed to modify the pump's diameter and height so that the modified pump could inflate a tire to a given pressure within a given time. The

pump simulation was designed to simulate the physics of a real pump as closely as possible.

The structure of the design task was intended to reflect a prototypical design problem: provide performance requirements that the design must satisfy, set constraints on the design, and allow participants to bring to bear any processes they believe appropriate to solving the problem. This general approach has been used in a variety of contexts: evaluating engineering education (Atman et al., 1999; Katz & James, 1998; Mullins et al., 1999), assessing design skills in engineering (e.g., Katz & James, 1998), and assessing problem solving (e.g., Herl et al., 1999; Schacter et al., 1999). Our pilot testing of the pump design task suggested that participants did use a generate-and-test strategy.

Our expectation was that performance of the simulation task would be facilitated by knowledge of how bicycle pumps operated. However, because the task was a design task with specific pump performance requirements, knowledge may be insufficient. That is, while previous research has found a strong relationship between knowledge of pumps and problem-solving performance (e.g., Herl et al., 1999; Mayer & Gallini, 1990), the problem-solving task in prior work typically required participants to diagnose faults or list possible ways to increase the efficiency of the pump. In our simulation task, participants were asked not only to redesign a pump to meet performance criteria, but they could test their design and receive feedback on the adequacy of their design. Thus, we expected participants to engage in the generate-and-test strategy because the simulation provided both a specific pump performance objective and the capability for revision.

Simulation Interface and Embedded Process Measures

The Virtual Interactive ITS Development Shell (VIVIDS) was the simulation environment used to develop and administer the design problem. VIVIDS is an authorable, multimedia simulation environment that can be used for either instructional or assessment purposes (Munro et al., 1997; Munro & Pizzini, 1998). In VIVIDS, simulations are built by specifying the behavioral rules among interacting components. These behavioral rules govern the operation of the (simulated) system and can be used to develop high-fidelity simulations of complex systems. The programming facilities within VIVIDS provided the capability to record all student interactions with the system as well as the states of all components.

Pilot Test

We informally pilot tested the usability of the simulation interface with 5 participants who were unfamiliar with the content but who were comfortable using Graphic User Interfaces. The pilot test yielded no user interface problems. That is, we did not observe directly nor did the pilot-test participants report any problems with unintentionally clicking on a box, understanding the meaning of the various interface elements, or the notion of clicking on an icon to view information.

Interface Areas

As shown in Figure 2, the interface was divided into six major areas: (a) the diagram of the pump, (b) the task information section, (c) the information definition section, (d) the pump control section, (e) the design section, and (f) the solve problem section.

Pump Diagram

The pump diagram was based on a previous study that involved pumps (Herl et al., 1999). The diagram was dynamic in that the piston and valves would operate when the pump simulation was run. The purpose for having a dynamic display was to visually convey the operation of the pump (e.g., pushing down the piston caused the outlet value to open).

Task Information

Task information was presented as a simple statement of what the participant was to do: modify the pump to meet 1) the target tire pressure and 2) the target time-to-inflation. The purpose of this information was to provide participants with the performance requirements for the simulation task. This information was static and visible throughout the task.

Definitions and Pump State Values

As shown in Figure 2, there are shaded boxes next to labels and shaded boxes next to open boxes. The open boxes displayed the critical values associated with the pump simulation. The labels above the open box described the contents of the box (e.g., volume under piston, elapsed pumping time). To see a definition of the information, participants could click on the box to the left of the label, and the definition would show up in the "DEFINITIONS" section. To see the actual value of each pump parameter, participants were required to click on the shaded box to the left of the open box, below the label in question, and hold the mouse down. Releasing the mouse would mask the value.

The purpose of masking the value information was to provide a way to measure what the participant intended to view. In the absence of think-aloud protocols or eye-gaze data – both impractical for feasibility reasons – we believed our approach would be a simple and effective way to measure what the participant intended to view. We assumed because there was a usability cost associated with clicking on the button, participants would only click on the button if they intended to view the information and they believed the information was relevant or valuable. All requests to view information were time stamped and logged.

Pump Controls

The pump was "run" or simulated using the controls in this section. Participants could run the pump with 1 stroke or 30 strokes. The Reset Pump button zeroed out the pump simulation values and allowed participants to start over. Every request to "run" was time stamped and logged.

Design Section

The design section provided participants with a way of modifying the pump diameter or height. Clicking on higher numbers increased the dimensions of the pump and clicking on lower numbers decreased the dimensions of the pump. Changing the dimensions of the pump reset the pump parameter values. Every request to modify the pump diameter or height was time stamped and logged.

Solve Problem Section

The solve problem section showed (a) the participant ID in large type size, (b) the current problem number out of a maximum of 3, and (c) for the current problem, the attempt number out of a maximum of 3. Figure 2 shows an example of a screen at startup – the participant ID is 1000, the problem number is 1 and the participant has not made an attempt to solve the problem.

When the participant clicked on the Solve Problem button the solution was evaluated given the pump diameter and height values selected in the design section. If the design values were adequate to meet the specifications shown in the task section, then a pop-up window would appear informing the participant that the problem was solved and the participant was advanced automatically to the next problem. If the design values were not adequate to meet the performance specifications, then a pop-up window would appear informing the solution was incorrect, the attempt number would increment by one, and the participant would be returned to the simulation task. If the participant attempted to solve a problem three times without reaching a solution, the participant would be informed that he or she did not solve the problem and would be advanced to the next problem. No other feedback was given to the participant.

Figure 2

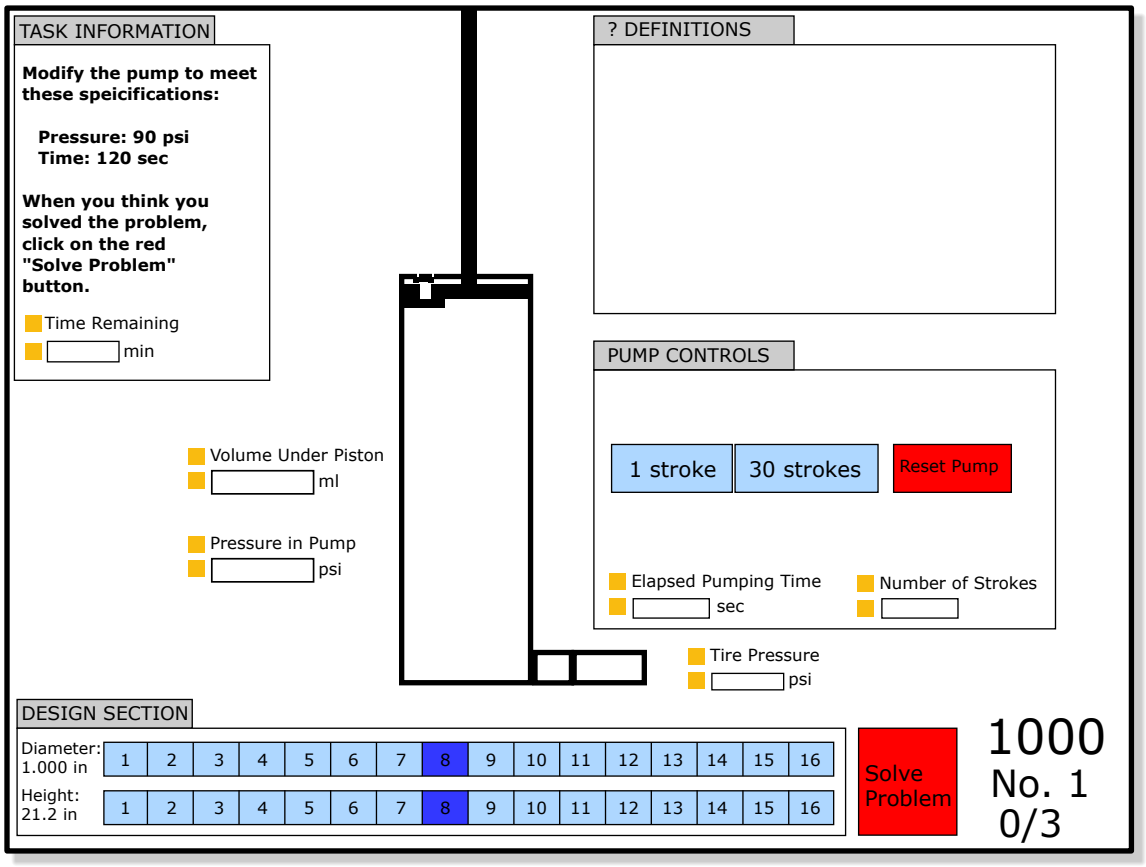

Figure 2. An example of the simulation screen image at startup.

Measures

Our overall measurement approach was to gather evidence of knowledge and reasoning skills expected to be related to the problem-solving process and performance on the simulation task. We expected to find positive relationships between performance on our scientific reasoning and two prior knowledge measures and performance on the simulation task. Such a finding would support the interpretation that our simulation task was operating as intended (i.e., as a design task). Based on existing literature, reasoning and knowledge of pumps were assumed to be important variables that would affect participants' problem-solving performance and processing (e.g., Baker & O'Neil, 2002; Baker & Mayer, 1999; Herl et al., 1999; Mayer & Wittrock, 1996; O'Neil, 1999). For example, in a prior study that used pumps and that was the basis for content used in the simulation, Herl et al. found a high correlation (r = .84) between scores of knowledge of pumps and performance on problem-solving problems requiring use of knowledge of pumps.

Background Information

The following information was collected from participants: age, ethnicity, sex, high school GPA, and SAT I verbal and math scores.

Familiarity with bicycle pumps and mechanical systems. To measure level of familiarity with bicycle pumps and mechanical systems, two 5-point Likert scale questions were administered to participants: (a) "In general, how familiar are you with how bicycle pumps work?", with anchor points described as 1 = "*not familiar*", 3 = *somewhat familiar*, and 5 = *very familiar – I already knew the concepts very well*. To measure background in mechanical devices, participants were asked the question: "How extensive is your background in mechanical devices (e.g., pumps, cars, etc.)?", with anchor points described as 1 = *little or no experience*, 3 = *somewhat extensive*, and 5 = *very extensive*.

Scientific Reasoning

Lawson's Classroom Test of Scientific Reasoning (CTSR) (revised 24-item multiple-choice edition) was used to measure scientific reasoning (Lawson, 1987). Two items (questions 21 and 22) were dropped because of a typographical omission. Coefficient alpha was .80 for this measure. The purpose for including the CTSR was to gather information on participants' reasoning, as we presumed the design task required reasoning to successfully solve it.

Prior Knowledge Measure of How Pumps Operate

An online knowledge map was administered to participants to measure their knowledge of how pumps work. This measure was administered before the simulation task and was intended to provide a baseline measure of students' understanding of pumps. The set of terms (i.e., *handle – up, handle – down, piston – up, piston – down, inlet valve – open, inlet valve – closed, cylinder – high pressure, cylinder – low pressure, outlet valve – open, outlet valve – closed, hose – airflow, hose – no airflow*) and links (i.e., *causes, contributes*) and the expert map used for scoring participants' knowledge maps were based on another study (Herl et al., 1999). The expert knowledge map was used as the criterion map for scoring purposes. The scoring algorithm was based on the method developed by Herl, Baker, and Niemi (1996). Briefly, the knowledge map score was the number of propositions in the participant map (i.e., concept-relationship-concept) that also existed in the criterion map. Because the student and expert maps were computer-based, the scoring was carried out automatically.

The rationale for measuring participants' pre-simulation knowledge was to gather information on participants' pre-simulation understanding of how bicycle pumps operate. Presumably, participants who knew more about the operation of pumps would be more likely to perform higher on the simulation task.

Prior Knowledge Measure of Pump Principles

A 12-item multiple-choice test was used to gather information on participants' understanding of the relationships between the design of a pump (i.e., its height and width) and stroke time, volume, pressure, and force. An understanding of these relationships was assumed to be important to successfully solve the pump design problems. We expected that participants who had higher knowledge of pump principles also would perform higher on the simulation task. Coefficient alpha was .81 for this measure.

Simulation Outcome Measures

The three measures of performance on the simulation task were the number of problems solved, the number of incorrect solve attempts, and the time taken to complete the task. Three simulation problems were developed to measure participants' ability to solve the pump design problems. Participants had a maximum of three chances to solve each problem. The first problem was designed to be the easiest, and the last problem the hardest. Difficulty was determined by the solution range. The more difficult problem had a much narrower range of solutions (i.e., fewer height-width combinations). An expert experienced in the design and operation of pumps designed the problems and solution ranges.

Simulation Online Process Measures

Seventeen events were used for analyses. These events represented the click-stream activity of participants and are given in Table 1. These events were grouped into four event categories: (a) design, (b) run, (c) information, and (d) solve attempt. The four categories were intended to reflect the general generate-and-test strategy framework. The adequacy of these categories was verified by expert review. The expert had extensive engineering experience with the design and operation of pumps.

Table 1 Simulation Events

Type of event	Description
Design	Clicked on a design button
Run	Clicked on 1 stroke button
	Clicked on 30 stroke button
	Clicked on pump reset button
Information	
Definition	Viewed time remaining definition
	Viewed volume under piston definition
	Viewed pressure under piston definition
	Viewed elapsed pumping time definition
	Viewed number of strokes definition
	Viewed tire pressure definition
Value	Viewed time remaining value
	Viewed volume under piston value
	Viewed pressure under piston value
	Viewed tire pressure value
	Viewed elapsed pumping time value
	Viewed number of strokes value
Solve attempt	Clicked on solve button

Impact of User Interface Measure

To measure participants' perceptions of the user interface, three 5-point Likert scale questions were administered to participants:

(a) "In general, how intrusive did you find having to click to see information?", 1 = *not intrusive*, 3 = *somewhat intrusive*, 5 = *very intrusive*;

(b) "How frequently did having to click on the information boxes interfere with your performance on the task?" 1 = *clicking did not interfere with my performance*, 3 = *clicking interfered with my performance sometimes*, 5 = *clicking interfered with my performance very often*;

(c) "Compared to not having to click to see information, how often did clicking for information change your thinking on the task?" 1 = *clicking for information did not change my thinking*, 3 = *clicking for information changed my thinking sometimes*, 5 = *clicking for information changed my thinking very often*.

In addition, participants were asked to respond to the following question in an open-ended response: "What effects, if any, did explicitly clicking have on you during the computer task? Please describe or check 'no effects' if there were no effects."

Procedure

Because participants had participated in a previous study, we had available the background and scientific reasoning test scores. For the current study, participants were first given a diagram of a bicycle pump, with the different parts of the pump identified. Next, the pump principles test was administered. Participants were given as much time to complete the 12-item survey as they needed. In general, they took about 2 minutes.

Next, participants were administered the knowledge map task. All participants were familiar with the knowledge map software as they had used it in a prior study. The knowledge map task was used to depict cause-effect relations among the different pump elements. Participants received instruction and examples on paper. These procedures were adopted from Herl et al. (1999). The training on how to use the knowledge mapping software took about 5 minutes and participants were given 10 minutes to complete the knowledge map.

Following the knowledge map task, participants received a diagram depicting how bicycle pumps work (Herl et al., 1999). Participants were then shown the simulation computer interface and given instructions on the task. Participants were given 45 minutes to complete the task which included three problems. Following the simulation task participants were given 3 minutes to modify their knowledge map, followed by a new pump principles test where they were given as much time as they needed to complete the test. In general, participants took about 2 minutes to complete the test.

Participants were then administered the familiarity with bicycle pumps and mechanical systems measure and the user-interface evaluation measure, debriefed, and paid for their time.

Results

The analyses of problem-solving patterns were based on participants' clickstream data. Participants' clickstream data were analyzed using lag sequential analyses. The analyses of perceived effects were based on participants' self-reports to survey questions. All statistical tests were two-tailed and the *p*-value set to .05.

Preliminary Analyses

Prior to analyzing the data, we conducted a preliminary analysis of the pattern of correlations among our measures to determine the degree to which our measures were operating as expected. Given the nature of the simulation task – a design problem presented to participants with little knowledge of how pumps operated – we expected to observe correlations among reasoning skills, pump knowledge, and performance on the task. As shown in Table 2, there were no significant relationships between the pretests of pump knowledge and performance variables. The pattern of correlations among the measures of pump knowledge and scientific

reasoning, and between these measures and the outcome measure were not significant; thus it is unclear whether there were problems with the measures or the simulation task did not require the degree of prior knowledge that we believed at first. Some evidence was found that the prior knowledge measures were operating as expected. A moderate correlation was found between the pretest pump operation test score and participants' self-reported experience with mechanical devices (r_{sp} = .35, p = .10, n = 23). We speculate that the simulation task did not require the depth of background knowledge we initially believed.

Table 2 Spearman Correlations Between Measures of Prior Knowledge and Scientific Reasoning and Task Performance, and Intercorrelations Among Measures of Prior Knowledge and Scientific Reasoning (N = 25)

				Simulation task performance		Prior knowledge measures	
		M	SD	No. of problems solved[a]	No. of incorrect solve attempts[b]	Pump principles	Pump operations
Prior knowledge measures	Pump principles[c]	6.76	1.79	.22	-.24	–	.17
	Pump operations[de]	2.70	1.82	.39	-.37	.17	–
CTSR[f]		12.84	4.54	-.04	-.04	.22	.17

a Maximum possible 3.
b Maximum possible 9.
c Multiple-choice measure, maximum possible 12.
d Knowledge map measure, maximum possible 12.
e n = 23.
f Multiple-choice measure, maximum possible 22.
* p < .05 (two-tailed).

Analyses of Participants' Problem-Solving Processes

Task Performance

Given 45 minutes to solve three problems, 2 participants solved no problems, 10 solved one problem, 5 solved two problems, and 8 solved all three problems. Participants' self-reported perceptions of the difficulty of the task are consistent with performance, as shown in Table 3.

Table 3 Number of Problems Solved by Perception of Task Difficulty ($N = 25$)

	In general, how difficult did you find the pump task?[a]				
No. of problems solved[b]	**1**	**2**	**3**	**4**	**5**
0			1		1
1			6	4	
2		1	1	1	2
3	1	2	4	1	

[a] 1 = *Not difficult*, 3 = *Somewhat difficult*, 5 = *Very difficult*.
[b] Maximum possible 3.

The higher participants' rating of task difficulty, the longer they took to complete the simulation ($r_{sp} = .56$, $p < .01$, $N = 25$). These results suggest that while the task may have been difficult for many participants, it was also easy to somewhat difficult for the majority of participants.

Table 4 shows the descriptive statistics and intercorrelations for the outcome and process measures. In terms of simple counts of behavior (i.e., the number of clicks), the most frequently used element of the simulation was the inspection of information followed by running the pump simulation and design activity. No other interesting relationships were found.

Table 4 Descriptive Statistics and Spearman Intercorrelations ($N = 25$)

		M	SD	Simulation outcome measures			Online process measures		
				No. of problems solved	No. of incorrect solve attempts	Task time	Design	Run	Information
Simulation outcome measures	No. of problems solved[a]	1.76	1.01						
	No. of incorrect solve attempts[b]	4.00	2.80	-.87**					
	Task time (seconds)[c]	1913	592	-.48*	.33				
Online process measures	Design	90.40	46.67	-.59**	.47**	.57**			
	Run	175.40	137.74	-.18	-.31	.30	.22		
	Information	235.08	99.84	-.22	.10	.71**	.62**	.38	
	Solve attempt	5.76	1.96	-.76**	.97**	.19	.36	-.37	.04

[a] Maximum possible 3.
[b] Maximum possible 9.
[c] Maximum possible 2700.
* $p < .05$ (two-tailed).
** $p < .01$ (two-tailed).

Online Processes

To examine participants' online problem-solving processes, we conducted a sequential analysis of participants' clickstream behavior with respect to the four main event categories listed in Table 1. A sequential analysis takes into account the order or sequence of behavior as it unfolds over time. We selected this type of analysis because our main objective was to measure problem-solving processes—participants' dynamic, moment-to-moment behavior over time (Bakeman & Gottman, 1997; Bakeman & Quera, 1995; Gottman & Roy, 1990). Sequential analyses have been used with a variety of data whose common elements are interaction and time (e.g., communication streams, Bowers, Jentsch, Salas, & Braun, 1998; Hirokawa, 1980; human-computer interaction, Sanderson & Fisher, 1994; marital interaction, Gottman, Markman, & Notarius, 1977).

Participants' behavior (as measured by the event categories in Table 1) was logged and the order of events preserved. For example, using the four major event categories (design, run, information, solve attempt), a hypothetical participant's clickstream might be "D R I S ..." This sequence of codes represents a participant clicking on a design button (D), followed by running the simulation (R), viewing a pump parameter (I), and making a solve attempt (S). Sequential analyses can quantify the time-dependent dependencies among the behaviors.

To determine the strength of relationship among the behaviors, 2 × 2 contingency tables were constructed for each pair of event types. For each pair of events,

the odds ratio was computed. For any 2 × 2 contingency table, rows represent the given event (given event occurred, did not occur) and columns represent the target event (target event followed given event, did not follow given event), as shown below in Figure 3.

Figure 3

		Target event	
		B	**not B**
Given event	**A**	a	b
	not A	c	d

Figure 3. Generic contigency table: strength of relationship between two behaviors, the given event (A) and the target event (B), with the cells (a, b, c, d) indicating the number of times that the given event occurring was followed by the target event occurring or not occurring.

Cell values represent the count of particular pairs of transitions (e.g., *a* is the number of times event B followed event A). The odds ratio is given as ad/bc and ranges from 0 to infinity. For ease of interpretation, Yule's Q was computed from the odds ratio. Yule's Q is the transformation of the odds ratio and is given as $[(ad - bc) / (ad + bc)]$ and ranges from –1 to +1, where –1 indicates perfect negative association, 0 no association, and +1 perfect positive association. Yule's Q is analogous to a correlation coefficient (Bakeman, McArthur, & Quera, 1996).

Table 5 shows Yule's Q for lag 1 and lag 2 sequences. The given event (lag 0) column specifies the behavior of interest, and the target event column specifies the event immediately following the given event (lag 1) or two events following the given event (lag 2). For example, given that a design event occurred, the likelihood of a run event following is very high (.89) while the likelihood of an information event following is very low (–.67).

Figure 4 depicts the order of events. Each node represents one of the major events (i.e., design, run, information, or solve attempt) and the arcs show Yule's Q between events. As Figure 4 shows, participants' behavior was cyclical – design, run, information, and then either design or solve attempt and then design. This behavior set was remarkably stable, with most lag 1 associations greater than .80.

Shown in Table 5 but omitted from Figure 4 are the negative associations. These associations indicate behavior that occurred far less than expected. For example, after running the simulation, participants rarely clicked on the design buttons or attempted to solve the problem. Likewise, design activity was rarely followed by looking at information.

Table 5 Yule's Q for Lag 1 and Lag 2 Sequences, All Participants (*N* = 25)

Given event (Lag 0)	Target event							
	Lag 1				Lag 2			
	Design	Run	Information	Solve attempt	Design	Run	Information	Solve attempt
Design		.89	-.67	-.38	-.67	-.79	.88	-.71
Run	-.75		.98	-.90	.83	.30	-.98	.77
Information	.89	.22		.82	-.70	.24	.30	-.65
Solve attempt	.59	-.42	-.22		-.13	.33	-.29	.02

Note. All values of Q are within a 95% confidence interval.

Figure 4

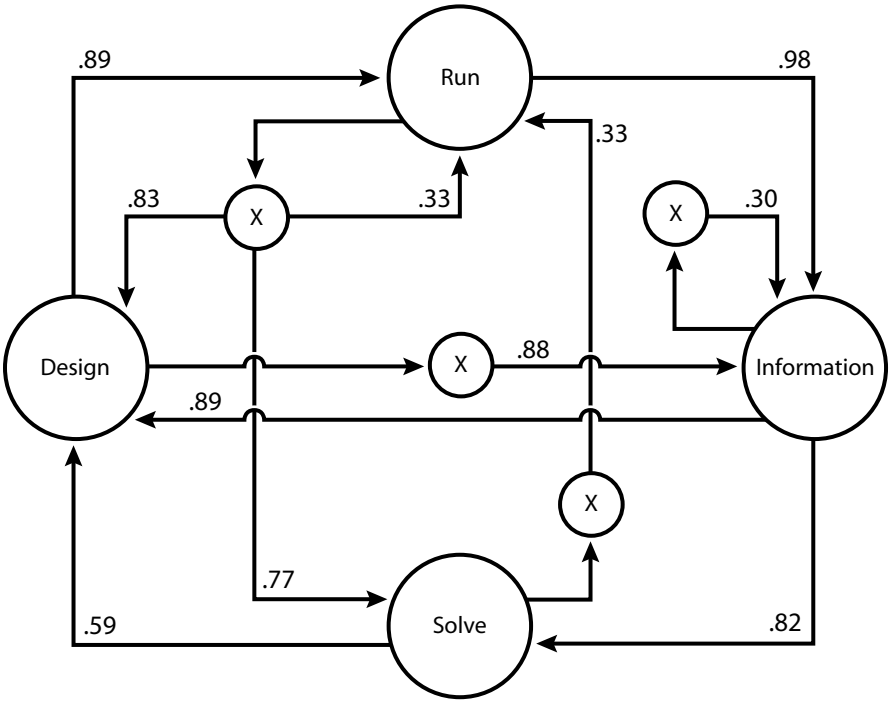

Figure 4. Lag 1 and lag 2 transitions (only values of Yule's Q greater than or equal to .30 are shown). All values are within a 95% confidence interval. X indicates any intervening event.

These results suggest that participants' behaviors were very systematic and consistent with the generate-and-test processes observed in previous studies using the think-aloud method (i.e., Katz & James, 1998; Nhouyvanisvong & Katz, 1998). In particular, we interpret participants' sequential behavior – their design activity (e.g., modifying the pump), followed by running the simulation, viewing pump parameter information, then attempting to solve the problem – as behavior indicating that a participant was generating a hypothesis, testing the hypothesis, and

then revising the hypothesis. The mapping between sequential behavior and the generate-and-test process is shown in Figure 5.

Figure 5

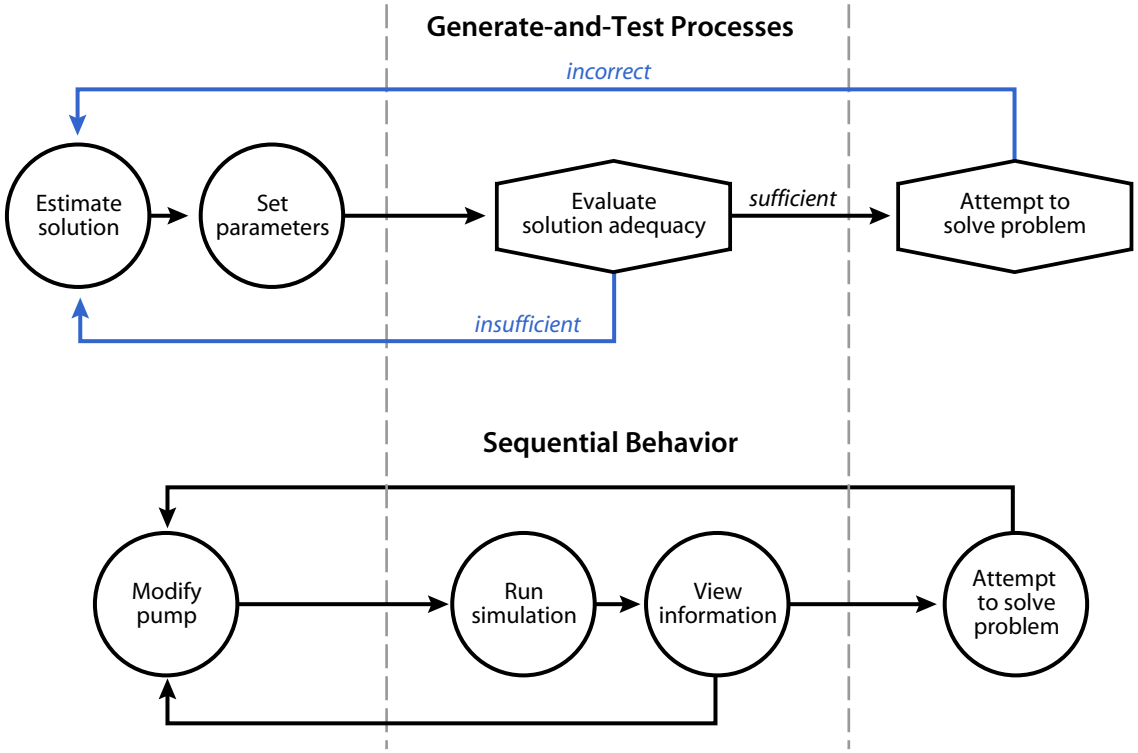

Figure 5. Mapping between generate-and-test processes and online sequential behavior

Our results suggest that user-interface events can be used as measures of problem-solving processes, particularly if the sequential behavior is taken into account. To elaborate, one aspect rarely addressed in studies of problem solving is the sequential (or time-dependent) dimension of problem solving. By definition, process implies time or sequence. Taking sequence into account yields results that describe how a participant's behavior unfolds over time. For example, the correlations among the online process measures show a significant relationship between design activity and information events (see Table 4); not much more can be inferred. However, the complementary information shown in Table 5 and Figure 4 describes the nature of the relationship – design events are less likely to be followed by information events than solve attempts. Further, this relationship is asymmetrical; both information and solve attempt events are very likely to be followed by design events. Finally, the sequential pattern shown in Figure 4 and summarized in Figure 5 captures the notion of process as a time-dependent phenomena that cannot be inferred from simple correlations as shown in Table 4.

Analyses of Interface Effects

To determine interface effects we asked participants their perceptions of the effects of the click-through interface. Table 6 shows the user interface evaluation questions and distribution of participants' responses. Most of the participants reported that the interface had some effect on them in terms of intrusiveness or interference on the task; 6 participants reported no effects at all, 3 reported the interface as being very intrusive, and 6 reported the interface to have interfered with task performance very frequently.

Table 6 User Interface Evaluation (*N* = 25)

	Response				
Question	**1**	**2**	**3**	**4**	**5**
In general, how intrusive did you find having to click to see information?[a]	6	3	12	2	2
How frequently did having to click on the information boxes interfere with your performance on the task?[b]	6	2	10	2	5
Compared to not having to click to see information, how often did clicking for information change your thinking on the task?[c]	5	4	6	8	2

a 1 = *Not intrusive*, 3 = *Somewhat intrusive*, 5 = *Very intrusive*.

b 1 = *Clicking did not interfere with my performance*, 3 = *Clicking interfered with my performance sometimes*, 5 = *Clicking interfered with my performance very often*.

c 1 = *Clicking for information did not change my thinking*, 3 = *Clicking for information changed my thinking sometimes*, 5 = *Clicking for information changed my thinking very often*.

When participants were asked if the interface changed their thinking on the task, the responses were evenly distributed. Of the 25 participants, 15 participants reported that the interface changed their thinking sometimes or less, while 2 reported that the interface changed their thinking very often. The remaining 8 participants reported the interface changed their thinking between sometimes and very often. Participants who reported the interface as intrusive also tended to report the interface as interfering with task performance ($r_{sp} = .62$, $p < .01$, $N = 25$) and had a higher perception of the difficulty of the task ($r_{sp} = .62$, $p < .01$, $N = 25$). In terms of performance, the more participants reported the interface as intrusive, the lower the number of problems they solved ($r_{sp} = -.58$, $p < .01$, $N = 25$), the more incorrect solve attempts they made ($r_{sp} = .46$, $p < .05$, $N = 25$), and the longer they took to complete the task ($r_{sp} = .42$, $p < .05$, $N = 25$). Interestingly, there were no significant relationships between participant reports of how much the user interface interfered *with their thinking* and any of the task performance measures, the self-reported intrusiveness of the user interface, or the self-reported intrusiveness of the interface on task performance.

When participants were asked to describe the effects the interface had on them, 9 participants reported no effects. The remainder of participants responded in writing, and their responses were classified into three categories – (a) slow down on task, (b) negative impact on the task performance, and (c) positive impact on task performance. Two responses were not classified.

Finally, when task performance and participants' self-reports of the interface are considered together, the pattern of correlations suggest a possible cognitive load effect (Sweller, 1994). That is, participants who reported the user interface as being intrusive also performed more poorly on the simulation task.

This finding, while tentative, is consistent with the general idea of the mediating role of prior knowledge on problem-solving performance (e.g., Mayer, 2001; O'Neil, 1999). We speculate that for participants with low prior knowledge, the joint effects of (a) a novel content area; (b) the task demand to determine the causal relationships among the pump's diameter, width, volume, and pressure; and (c) an unusual interface that taxed working memory by hiding critical status information, may have imposed too severe a processing load on low-knowledge participants. However, note that we are speculating about the presence of a cognitive load effect, as we did not observe any significant relationships between our prior knowledge measures and task performance measures.

Discussion

This study examined a novel user-interface technique to explicitly measure participants' problem-solving processes. We found strong evidence of a sequential structure to participants' behavior, consistent with a generate-and-test strategy. Participants' problem-solving processes were strikingly consistent: (a) design activity, (b) run the simulation, (c) check the information, and (d) back to design or solve attempt and then back to design.

Limitations and Next Steps of This Work

Our approach of requiring participants to explicitly click on a button to view important information appeared successful; however, over half of the sample reported that the interface was intrusive or it interfered with the task in some way. Despite these self-reports, participants were able to engage in the task and there was no relationship between participants self-reports of the frequency of the interface changing their thinking, and the measures of intrusiveness of the interface or task performance. Thus, while participants' reported the interface as intrusive, it is unclear to what extent their performance was affected by the interface. Future research should investigate this issue in more depth (Bennett & Bejar, 1998).

A general limitation of this work is that it is confined to describing the problem-solving processes of a particular sample using a particular online task. The study lacked a comparison group or groups. While the results of this study suggest

that the general approach is promising, future research should address the following issues:

- Compare performance on the click-through interface to performance on a non-click-through interface. One outstanding issue is to what extent does the interface interfere with task performance. Such a condition can provide information on the impact of the click-through interface on task performance (i.e., number of problems solved; number of solve attempts), efficiency (performance per unit time), and click rate (number of clicks on both the design and simulation buttons).
- Comparisons of the sequential aspects of participants' problem-solving processes should be compared for different sub-groups. For example, three needed comparisons are between (a) low- and high-domain knowledge participants, (b) low- and high-task-performing participants, and (c) low- and high-reasoning participants. These comparisons flow directly from theoretical considerations of the factors that influence problem-solving performance (see Mayer and Wittrock [1996] and O'Neil [1999] for a discussion). Detection of problem-solving process differences, as measured by the sequence of behaviors, between participants in the various conditions would bolster the argument that our approach supports the measurement of problem-solving processes.

Implications for Online Assessment

The significance of this work lies in three areas. First, we have tested a low-cost methodology to capture participants' overt behavior that went well beyond capturing arbitrary keystrokes and clicks. The technique we used was to make obvious the availability of important information and elements of the task, but provide that access at a cost – participants had to explicitly click on a button. Our reasoning was that participants would not engage in such (costly) activity unless they perceived a need for the information; thus, we assumed participant clicks represented cognitively meaningful behavior. Current means of capturing process data are expensive in terms of time and effort (e.g., behavioral observations, think-aloud protocols).

The second implication of the work is that we have empirical support, although exploratory, of the quantification of problem-solving processes from clickstream data, and the corresponding statistical support for the intuitive notion that sequential dependencies exist in participants' problem-solving strategies. We adopted a methodology that is routinely used in behavioral observation research (e.g., see Bakeman & Gottman, 1997) and have applied it to characterize online problem-solving behavior. The utility of this approach is that it allows us to examine the processes underlying performance. This study represents an early step toward developing a methodology and analysis approach that can take advantage of clickstream data.

Finally, the third area this study contributes to is that when considered in the context of prior work (e.g., Chung et al., 2002; Chung, Kim, de Vries, & Phan, 2003; O'Neil, Chuang, & Chung, in press), the current study provides another

example of how clickstream data can be used to infer cognitive processes. The current study followed up on the idea first presented in Chung et al. (2002) that meaningful information can be extracted from the user interface to the extent that the interface supports the unambiguous capturing of intentional acts. In the current study, how to run the simulation was visible and obvious to the participant, and its use was inferred as an intentional act. Similarly, the information presented to the participant was unambiguous. Each piece of information could only be viewed via a mouse click and only a single piece of information was presented at a time (vs. multiple pieces of information); thus, we assumed there was little doubt about what the participant was viewing. These are important features because presumably, the act of clicking reflects the result of participants' reasoning and judgment.

Implications for Online Instruction

One of the most promising aspects of online delivery of tasks is that the information gathered about the student from the student's interaction with the task itself has the potential to individualize instruction. While this idea is not new, what is new is that our understanding of what to do with the data is much more sophisticated. For example, when interpreting the meaning of the click, the usefulness of the click – the reasonable inferences that can be drawn – is increased substantially when the set of cognitive demands behind the click is clear (e.g., Chung, et al., 2002, 2003; O'Neil et al., in press). Thus, one immediate instructional application of the general approach tested in this study would be to compare individual student processes against one or more criterion processes, and provide individualized feedback to the student. For example, if a task has a range of sequences that could be characterized in terms of degree of optimality or correctness, then feedback on the extent an individual student's problem-solving process compared to a criterion process could easily be administered.

A more long-term application of higher instructional utility would be to use the clickstream data as a way to link assessment and instruction *in real-time*. That is, inferences about student problem-solving processes could be summarized across students or reported at the individual student level and provided back to the instructor *as the student engages in the online task*. We are currently pilot testing the utility of such feedback for instructors, although the feedback on student performance is in the form of individual student typed responses (via an anonymous chat-like interface) to electrical engineering circuit problems (Ainsworth, Guidry, Chung, Delacruz, & Kaiser, 2003). One key insight of this work is that student process data obtained via the anonymous chat interface – in the form of intermediate solutions to problems and questions to the instructor – provide far more useful information (compared to a traditional face-to-face classroom setting), and this information can be acted on instructionally in real-time. We expect a similar kind of utility from clickstream data to the extent the clickstream data accurately reflect the outcome of substantial cognitive processing.

As more problem-solving tasks are delivered online, the use of clickstream data will become increasingly important as one source of data. Online problem-solving

tasks are essentially software driven. Consequently, embedding measures of problem-solving processes in the task becomes a matter of defining and programming appropriate software sensors. However, our confidence in the interpretation of student telemetry is improved markedly when the clickstream data accurately reflect the participant's intention and are the outcome of meaningful cognitive events.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Ainsworth, E., Guidry, T., Chung, G. K. W. K., Delacruz, G. C., & Kaiser, W. L. (2003, April). *UCLA/Hewlett-Packard Computer Science/Engineering Retention Project* (Quarterly Report). Los Angeles: University of California, Center for Excellence in Engineering & Diversity.

Atman, C. J., Chimka, J. R., Bursic, K. M., & Nachtmann, H. L. (1999). A comparison of freshman and senior engineering design processes. *Design Studies, 20*, 131–152.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction* (2nd ed.). New York: Cambridge University Press.

Bakeman, R., McArthur, D., & Quera, V. (1996). Detecting group differences in sequential association using sampled permutations: Log odds, kappa, and phi compared. *Behavior Research Methods, Instruments, & Computers, 28*, 446–457.

Bakeman, R., & Quera, V. (1995). *Analyzing interaction*. New York: Cambridge University Press.

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior, 15*, 269–282.

Baker, E. L., & O'Neil, H. F., Jr. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior, 18*, 609-622.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*(3), 5–12.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement, 17*(4), 9–17.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors, 40*, 672–679.

Chung, G. K. W. K., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior, 18*, 669–684.

Chung, G. K. W. K., Kim, J.-O., de Vries, L. F., & Phan, C. H. (2003). *Validating a method to infer learning processes from users' clickstream data using Bayesian networks*. Unpublished manuscript.

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. E.

Burstein (Eds.), *Automated essay grading: A cross-disciplinary approach* (pp. 23–40). Mahwah, NJ: Erlbaum.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*, 310–324.

Gottman, R. M., Markman, H., & Notarius, C. (1977). The topology of marital conflict: A sequential analysis of verbal and nonverbal behavior. *Journal of Marriage and the Family, 39*, 461–477.

Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers*. New York: Cambridge University Press.

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of U.S. history knowledge. *Journal of Educational Research, 89*, 206–218.

Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., Bianchi, C., Wang, S.L., Mayer, R. E., et al. (1999). *Final report for validation of problem-solving measures*. (CSE Tech. Rep. No. 501). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hirokawa, R. Y. (1980). A comparative analysis of communication patterns within effective and ineffective decision-making groups. *Communication Monographs, 47*, 312–321.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based scoring. *Educational Measurement: Issues and Practice, 20*(3), 16–25.

Katz, I. R., & James, C. M. (1998). *Toward assessment of design skill in engineering* (GRE Research Report 97–16). Princeton, NJ: Educational Testing Service.

Lawson, A. E. (1987). *Classroom test of scientific reasoning: Revised paper-pencil edition*. Tempe: Arizona State University.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E., & Gallini, J. (1990). When is an illustration worth ten-thousand words? *Journal of Educational Psychology, 82*, 715–726.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York: Macmillian Library Reference USA, Simon & Schuster Macmillan.

Mullins, C. A., Atman, C. J., & Shuman, L. J. (1999). Freshman engineers' performance when solving design problems. *IEEE Transactions on Education, 42*, 281–287.

Munro, A., Johnson, M., Pizzini, Q., Surmon, D., Towne, D., & Wogulis, J. (1997). Authoring simulation-centered tutors with RIDES. *International Journal of Artificial Intelligence in Education, 8*, 284–316.

Munro, A., & Pizzini, Q. A. (1998). *VIVIDS reference manual*. Los Angeles: University of Southern California, Behavioral Technology Laboratories.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment, Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). [Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education.] Washington, DC: National Academy Press.

Nhouyvanisvong, A., & Katz, I. R. (1998). The structure of generate-and-test in algebra problem-solving. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 770–775). Hillsdale, NJ: Erlbaum.

O'Neil, H. F., Jr. (1999). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior, 15*, 225–268.

O'Neil, H. F., Jr., Chuang, S. S., & Chung, G. K. W. K. (in press). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education*.

Sanderson, P. M., & Fisher, C. (1994). Exploratory sequential data analysis: Foundations. *Human-Computer Interaction, 9*, 251–317.

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., & O'Neil, H. F., Jr. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior, 15*, 403–418.

Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, J., & Clyman, S. (1999). Artificial neural network-based performance assessments. *Computers in Human Behavior, 15*, 295–313.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312.

Towne, D. M., Munro, A., Pizzini, Q. A., Surmon, D. S., Coller, L. D., & Wogulis, J. L. (1990). Model-building tools for simulation-based training. *Interactive Learning Environments, 1*, 33–35.

Authors' Note

The work reported herein was supported by a grant from GTE/Verizon as administered by the Center for Digital Innovation at the University of California, Los Angeles, and the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of Verizon, the Center for Digital Innovation/UCLA, the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

We would like to thank John Propst of UCLA for programming the simulation and deriving the equations necessary to simulate the pump operations. We would

also like to thank Dr. Allen Munro of the Behavioral Technology Lab, University of Southern California, for use of the VIVIDS system, and Dr. Anton Lawson of the University of Arizona for the use of the Classroom Test of Scientific Reasoning, and three anonymous reviewers for their helpful comments. Finally, we would like to thank Joanne Michiuye of UCLA/CRESST for editorial help with this manuscript.

Correspondence concerning this article should be addressed to Gregory K. W. K. Chung, UCLA CSE/CRESST, 10945 Le Conte, 1400C, Box 957150, Los Angeles, California 90095-7150. e-mail: greg@ucla.edu.

Author Biographies

Gregory K. W. K. Chung is a Senior Research Associate at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). His current work at CRESST involves developing problem-solving assessments for computer-based assessments and developing Internet-based assessment tools for diagnostic and embedded assessment purposes. Dr. Chung earned a Ph.D. in Educational Psychology from the University of California at Los Angeles, an M.S. degree in Educational Technology from Pepperdine University at Los Angeles, and a B.S. degree in Electrical Engineering from the University of Hawaii at Manoa.

Eva L. Baker is co-director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) and director of the UCLA Center for the Study of Evaluation. Her research has long focused on the integration of teaching and measurement, including design of instructional systems and new measures of complex human performance. She has conducted studies of technology, teaching, learning, and assessment in history, science, workforce readiness, and writing. Dr. Baker's current focus is on the uses of technology and performance assessment in large-scale environments for military training and public education. She has conducted research for or advised the U.S. Congress, the National Education Goals Panel, the Office of Technology Assessment, the Departments of Education, Labor, Energy, and Defense, the National Business Roundtable, private foundations, state departments of education, local school districts, and private corporations. She has advised ministries and universities in Latin America, the Middle East, Australia, Europe, and Asia, and international organizations such as NATO. She is a member of the U.S. Department of Education Advisory Council on Education Statistics, the Independent Review Panel on Title I, and was the measurement expert in the Title I negotiated rule-making process. Dr. Baker has more than 450 publications.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College