Cognitive Validation of a Computer-Based Assessment of Problem Solving: Linking Cognitive and Online Processes

Gregory K. W. K. Chung[1], Linda F. de Vries[1], Alicia M. Cheak[1], Ronald H. Stevens[2], & William L. Bewley[1]

[1]National Center for Research on Evaluation, Standards, and Student Testing (CRESST) / [2]UCLA School of Medicine / [2]Graduate School of Education & Information Studies

Abstract

In this study we tested a novel cognitive validation strategy that yoked participants' verbal protocols with their clickstream data using a problem solving assessment (IMMEX—Interactive MultiMedia EXercises). Participants were presented with a scenario and provided with relevant and irrelevant information to solve the task. Participants could access the information in any order and attempt to solve the problem at any time. The most frequently occurring cognitive processes were paraphrasing text, making accurate cause-effect inferences, and monitoring of problem solving behavior. Productive processes were related to success and consistent with scientific reasoning behavior on the task and unproductive processes were related to unsuccessful performance and consistent with poor reasoning. We found strong evidence of the cognitive validity of the IMMEX task. Components of an online process-based assessment testbed are identified.

Embedded Assessment of Complex Skills

Engineering education has been a leader in the novel use of computer-based instruction. The intended use of many of the applications is to increase students' understanding of the content and to develop their problem solving skills in a particular content area. This is consistent with recent calls for engineering schools to increase students' problem solving skills[1-10]. However, assessing cognitively complex skills such as problem solving requires use of modern assessment methods; typical course evaluations, anecdotal instructor evaluations, and surveys of student attitude are inadequate[1-4].

One method being explored is the use of computer-based instructional applications as assessment platforms. Essentially, the idea is to provide students with instructional tools and embed assessments of complex learning and problem solving in the tool. From the students' perspective, the task appears instructional; however, embedded within the software are assessments of student performance[11-15]. This is a departure from most computer-based assessments, which are usually stand-alone assessments[16-24].

While this approach appears intuitively obvious, using instructional applications for assessment purposes requires that the application meet validity criteria[25-31]. Further, the stringency of the validity criteria varies depending on the stakes involved in the assessment results. Some examples of how assessment results are used include diagnostics, course credit, certification, and selection. For the purposes of assessing students' complex learning and problem solving, one of the most important validity criteria is that the task demand of participants the intended cognitive processes. That is, if claims are made that instruction using a particular computer-based task will improve students' problem solving, then evidence of students using the intended problem solving processes is essential.

Computer-Based Performance Assessments

Computer-based performance assessments designed to measure the skills and competencies of examinees on complex tasks can potentially broaden the range of skills assessed while simultaneously increasing the precision of measurement[11, 27-29, 31-34]. One direction computer-based assessments have taken is the use of complex open-ended constructed-response tasks. Examples include tasks that measure conceptual understanding[12, 21, 35], writing[17, 24], mathematical reasoning[16], Web search skills[14, 15], teamwork skills[18, 22, 23], patient diagnosis[20], and problem solving skills[36-38].

Much of the work on computer-based performance assessments has focused on feasibility issues; much less work has been done on exploring the cognitive demands of such tasks. By cognitive demands, we mean the set of cognitive processes that are required of examinees by the task. Cognitive complexity is a key validity requirement and underlies modern conceptions of validity, assessment design, and assessment validation[25, 30, 31, 33, 34].

The current study attempted to address this issue by testing a cognitive validation methodology. Our strategy was to focus on an existing, well-understood, and tested task. We used IMMEX (Interactive MultiMedia EXercises), which is a sophisticated computer-based assessment that has been used for instructional and assessment purposes[37, 44].

Problem solving in IMMEX. The general problem solving framework of an IMMEX task is to first present students with a problem scenario and then to provide students access to information that may or may not be useful to solving the problem. One problem solving strategy required by some IMMEX tasks is elimination[37]. Successful participants need to use data to eliminate candidate solutions to a problem. Effective use of this strategy is dependent presumably on scientific reasoning. Key cognitive processes underlying the use of the elimination strategy are (a) the interpretation of the available information in the context of other information and the problem scenario, (b) the identification of relevant information, and (c) the use of data to draw appropriate causal inferences and conclusions. We define problem solving as the process participants use to solve a problem when the solution is not immediately apparent to the problem solver[48]. Further, the elimination strategies and their underlying cognitive processes make up the domain-independent strategies. The domain-dependent strategies, which are not the focus of this research, would be the particular strategies used by participants due to the peculiarities of the content.

In the particular IMMEX problem we examined, True Roots, the scenario for the examinee is to use a variety of medical tests (e.g., blood typing) to determine the birth parents of Leucine (a character in the task). True Roots was developed by a team of biology teachers and its content was validated by UCLA genetics professors. Our research questions were:

1. What are the kinds of cognitive processing participants engage in while solving the True Roots IMMEX problem?

2. What is the relationship among participants' cognitive processes and their online behaviors and task performance?

METHOD

Participants

Eighty-nine incoming freshman participants were recruited from a major public university in southern California. Participants were first quarter freshmen and were paid for completing the study. We were interested in students closest to the high school population as the IMMEX task was designed for high school students.

Of the original 89 participants, six were dropped from the analyses because of equipment failure. There were 33 males and 49 females overall, and the ethnic distribution was as follows: 34 Asian American, 26 White, 8 Latino, 6 Biracial, 1 African American, and 7 unreported. In terms of achievement, self-reported high school mean GPA was 4.01 (*SD* = 0.31), mean SAT I Math was 669 (*SD* = 78) and Verbal was 631 (*SD* = 77). Participants reported moderate familiarity with the task content (*M* = 2.96, *SD* = 1.07, *n* = 79; 1 = *not familiar*, 5 = *very familiar*) and some background in the content (*M* = 2.43, *SD* = 1.01, *n* = 79; 1 = *little or no experience*, 5 = *very experienced*).

Design

Our focus on cognitive processing led us to use the think-aloud methodology[49]. We established two conditions, a think-aloud condition and a non-think-aloud (control) condition. We were concerned about potential reactive effects of talking aloud during the task. Fifty-four participants were randomly assigned to the experimental condition and 29 participants to the control condition.

Online Tasks

IMMEX training and assessment task. Participants practiced for five minutes on an IMMEX task to familiarize themselves with the user interface and task structure. The training task was identical in problem structure but much simpler. The assessment task, True Roots, required participants to apply their knowledge of genetics to solve a problem. True Roots involved a main character, Leucine, who suspected that she may have been switched at birth at the hospital; thus, she was in search of her biological parents. Participants were asked to assume the role of her friend and to help Leucine determine among five sets of parents the identity of her true parents. Participants had access to a variety of lab test procedures and information resources (i.e., non-data information). Each lab test would, if understood and applied properly, eliminate one or two

sets of parents from the pool of potential parents. The students also had access to worksheets to record their data. If participants solved the problem correctly, they proceeded to the next problem (which contained the same scenario with different data and a different solution) until all five problems were completed.

Measures

Because we assumed that successful IMMEX performance required reasoning with content, we administered a set of measures intended to gather information on participants' prior knowledge of the content and their reasoning skills. To measure prior knowledge, we developed a measure based on the major concepts covered in True Roots. To measure reasoning, we selected validated measures of scientific reasoning, inferential reasoning, and syllogistic reasoning.

Prior knowledge. A 20-item prior knowledge short-answer measure was developed based on the content of the True Roots task. The format and instructions of the task were based on the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) assessment development model[45]. All major topics in the True Roots task were covered by at least one item. Participants were required to provide a definition or state the significance of the term. Each item was scored on a 4-point scale, where 0 was awarded for no answer or for a response that indicated the participant missed the point and a 3 was awarded for a response that indicated high understanding of the term.

Reasoning measures. We administered three reasoning measures. Lawson's Classroom Test of Scientific Reasoning[50] (revised 24-item multiple choice edition) was used to measure scientific reasoning, Part I of the nonsense syllogisms test of reasoning[51] was used to measure participants' ability to tell whether correct conclusions were drawn from given statements, and Part I of the inference test of reasoning[51] was used to measure participants' ability to tell which conclusion could be inferred from given statements. All items for all reasoning measures were multiple choice.

Background information. A 24-item student survey was used to gather demographic information, high school biology course-taking history, SAT I verbal and math scores, self-reported familiarity with the content and ease of the task, and ease and completeness of the think-aloud for participants in the experimental condition.

IMMEX online measures. Online outcome and process measures were gathered from the IMMEX task. Table 1 lists the measures and their operationalization. The IMMEX online measures were computed as proportions. We reasoned that given the open-ended nature of the task, a proportion measure would best account for variation across individuals with respect to the distribution of a participant's behavioral processing.

Cognitive processing measures. We identified 18 indicators of cognitive processing which we believed distinguished high performers from low performers. However, due to low usage, only 12 indicators were retained.

Participants' verbal protocols were segmented by event and coded by two raters for the presence/absence of any of the processes shown in Table 2. Disagreements were resolved by consensus.

Table 1. IMMEX Online Measures

Measure	Definition
Outcome	
Percent solved	The percent of cases solved. There was a maximum of five cases. The percent solved varied depending on the number of cases attempted. Percent solved = (no. solved / no. attempted).
No. of successful solve attempts	Number of times a solve attempt was successful.
No. of unsuccessful solve attempts	Number of times a solve attempt was unsuccessful.
Process[a,b]:	
redundant	Proportion of tests that were redundant with respect to ruling out a parent. A redundant test was defined as the participant accessing a test that (a) provides sufficient information to eliminate a particular parent, AND (b) the particular parent could have been eliminated by a prior test.
data/lab tests	Proportion of tests that provided data (e.g., blood typing).
expert	Proportion of tests that provided "expert" opinion. In the True Roots problem, experts provided conjecture or opinion.
dictionary	Proportion of tests that were from the dictionary. The dictionary provided definitions of terms encountered in the text.
library	Proportion of tests that were from the library. The library provided explanations of the different tests.

[a]All proportion measures were computed with respect to the total number of events or pages accessed. [b]The term *test* refers to accessing a page in the IMMEX program. A page could contain data from a test, or it could be information (e.g., library). We retain the term *test* for convenience.

Table 2. Categories for Event Scoring

Cognitive process	Definition
Paraphrasing or echoing text	Participant states information only, not a cause-effect relationship. Little or no reason or explanation is given.
Accurate cause-effect inferences	Participant attempts to explain what is going on rather than merely state what is happening and arrives at a correct conclusion.
Inaccurate cause-effect inferences	Participant attempts to explain what is going on rather than merely state what is happening but arrives at an incorrect conclusion.
Evaluating information	Participant evaluates the validity of a relation or the content without making inferences. Can be correct, incorrect, or indeterminate.
Clarifying gaps in knowledge	Participant attempts to find answers to questions. This usually involves library or dictionary referencing.
Judgments of information relevancy	Participant differentiates between relevant and non-relevant information. Selective in the steps taken to solve the problem and accesses information that is most useful.
Confusion	Participant asks questions or makes statements which reflect confusion about the content.
Awareness of failure to understand	Awareness of failure to understand individual meaning of content.
Verifying information	Participant double checks understanding of conclusion.
Monitoring problem solving	Participant is aware of task goals and one's progress toward the goals.

An event was defined as an uninterrupted clickstream within the same family of tests, as shown in Table 3. Table 3 shows the clickstream that would result in three different events—fingerprint, pedigree, and birth certificates. Synchronizing the clickstream with verbal protocols allowed us

to yoke participants' online behavior to their cognitive processing. As with the online measures, the cognitive processing measures were computed as proportions.

Table 3. Example of an Event, Clickstream, and Protocol

Event	Clickstream	Participant's verbal protocol
1	fingerprint/leucine	... Fingerprints. Leucine. Arch... Watsons... The mom has a loop, the dad has an arch... Fingerprints. Leucine, Leucine... R times r. This is all...
1	fingerprint/watson	
1	fingerprint/cayetano	
2	pedigree/ikeda	Ikeda's family... Ikeda's family.
3	birth certificate/leucine	... Boy, girl, girl.

Procedure

Participants were assigned randomly to conditions. Participants in the think-aloud condition were tested one at a time, and participants in the control condition were tested two at a time. Participants were first introduced to the study, signed consent forms, and were administered three reasoning measures and the prior knowledge measure. Participants in the think-aloud condition were then trained in talking aloud while participants in the control condition advanced to the training task. The training task was followed by the assessment task. After the assessment task, participants were administered the background survey. The entire session lasted up to 105 minutes.

RESULTS

The first set of analyses examined the kinds of cognitive processing participants used while engaged in the True Roots task, and the second set of analyses examined the relationship among the cognitive processes, behavioral online measures, and the task outcome measures.

Manipulation check. Prior to conducting analyses, we examined the IMMEX performance data for condition differences between the think-aloud and control conditions. We were concerned about potential reactive effects of the think-aloud procedure. Participants in the think-aloud condition were asked of their perception of the ease of talking aloud (M = 3.75, SD = 1.29, 1 = *not easy*, 5 = *very easy*) and completeness of talking aloud (M = 3.73, SD = 0.94, 1 = *not complete*, 5 = *very complete*). Participants in the think-aloud condition reported slightly higher perception of task difficulty (M = 2.25, SD = 1.08, n = 52; 1 = *not difficult*, 5 = *very difficult*) compared to the control condition (M = 1.89, SD = 0.70, n = 27), although this difference was not significant. With respect to other variables, there was significant difference on the online process variable (number of successful solve attempts) favoring the control condition, $t(79.83) = 3.53$, $p < .001$, and a difference approaching significance on the proportion of tests that accessed the prolog, $t(1.87)$, $p < .07$. Separate analyses using neural network classification showed gender effects within each condition. Thus, in subsequent analyses we include data only from the think-aloud condition and acknowledge that the data may not generalize to a non-think-aloud setting.

What are the kinds of cognitive processing students engage in while solving IMMEX problems? To answer this question, we conducted several analyses of participants' cognitive processing as they engaged in the IMMEX task, their online behaviors, and their post-task self-reports.

Analyses were conducted on participants' think-aloud protocols to determine the specific kinds of processing participants used while they engaged in the task. Table 4 shows descriptive statistics and correlations among the 12 cognitive processing variables. As shown in Table 4, 94% of participants' cognitive processing was accounted for by the 12 processes. The three most frequently occurring processes were paraphrasing or echoing text, making accurate cause-effect inferences, and monitoring problem solving behavior.

Table 4. Descriptive Statistics and Intercorrelations (Spearman) of Participants' Cognitive Processes (n = 46)

	Cognitive process[a]	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1.	Paraphrasing or echoing text	.18	.07	–										
2.	Accurate cause-effect inferences	.13	.06	-.38*	–									
3.	Inaccurate cause-effect inferences	.03	.02	-.05	.14	–								
4.	Accurate evaluation of information	.09	.06	-.19	.71**	-.01	–							
5.	Indeterminate evaluation of information	.04	.03	-.15	.44**	.15	-.10	–						
6.	Inaccurate evaluation of information	.04	.03	-.11	-.01	.44**	-.39**	.29	–					
7.	Clarifying gaps in knowledge	.06	.05	.36*	-.50**	-.38*	-.30*	-.34*	-.21	–				
8.	Judgments of information relevancy	.07	.04	-.43**	-.19	.00	-.21	-.08	.08	-.22	–			
9.	Confusion with content	.09	.06	.10	-.56**	-.22	-.45**	-.20	-.08	.47**	-.02	–		
10.	Awareness of failure to understand	.04	.03	.00	-.37*	-.38**	-.33*	-.16	-.03	.29	.17	.17	–	
11.	Verifying information	.06	.04	-.11	-.07	.09	-.02	-.07	.16	-.24	.06	-.29*	-.19	–
12.	Monitoring problem solving behavior	.11	.05	-.26	.20	.19	.02	.20	.12	-.37*	.14	-.42**	-.12	.12

[a]Proportion measure.
*p < .05 (two-tailed). **p < .01 (two-tailed).

An interesting set of relationships was observed between the cognitive processing and background and reasoning measures. Positive relationships were found between cognitive processes that suggest understanding and measures of prior knowledge and reasoning. For example, participants' use of accurate cause-effect inferences and accurate evaluation of information were positively and significantly related to prior knowledge (r_{sp} = .39, p < .01; r_{sp} = .35, p < .05), scientific reasoning (r_{sp} = .38, p < .05; r_{sp} = .36, p < .05), and inferential reasoning (r_{sp} = .31, p < .05; r_{sp} = .33, p < .05). The mean score of the scientific reasoning measure was 16.02 (*SD* = 3.47, max. = 22) and the mean score of the inferential reasoning measure was 6.46 (*SD* = 1.82, max. = 10).

Conversely, negative relationships were found between cognitive processes that suggest shallow processing or lack of understanding and measures of prior knowledge and reasoning. For example, paraphrasing or echoing text were negatively related to scientific reasoning (r_{sp} = -.31, $p < .05$) and inferential reasoning (r_{sp} = -.29, $p < .05$). Clarifying gaps in knowledge was negatively related to scientific reasoning (r_{sp} = -.32, $p < .05$) and confusion with content was negatively correlated with the SAT I verbal measure (r_{sp} = -.47, $p < .01$) and syllogistic reasoning (r_{sp} = -.35, $p < .05$). The mean score for syllogistic reasoning measure was 8.72 (SD = 2.54, max. = 10). These relationships are consistent with the interpretation that our coding scheme captured processing related to reasoning.

As shown in Table 5, 87% of participants' online processing was accounted for by the five online processes. The most frequent test participants conducted was lab tests, accounting for 72% of participants' online behavior. Also shown in Table 5 is the percent solved. In general, participants were successful on the IMMEX task, solving four out of five cases. Interestingly, the number of successful solve attempts was negatively and significantly related to participants' use of the experts, library, and dictionary.

Table 5. Descriptive Statistics and Intercorrelations (Spearman) for IMMEX Online Measures (n = 46)

Measure	*M*	*SD*	1	2	3	4	5	6	7
1. Percent solved	.79	.29	–						
2. No. of successful solve attempts	3.80	1.56	.86**	–					
3. No. of unsuccessful solve attempts	2.91	2.46	-.47**	-.45**	–				
4. Prop. of tests that were redundant	.09	.07	-.20	-.27	.36*	–			
5. Prop. of tests that were lab/data related	.72	.08	.28	.42**	-.24	-.00	–		
6. Prop. of tests that were experts	.08	.04	-.21	-.31*	.01	.08	-.65**	–	
7. Prop. of tests that were library	.05	.04	-.17	-.37*	-.08	.14	-.58**	.24	–
8. Prop. of tests that were dictionary	.02	.02	-.04	-.16	-.04	.09	-.54**	.19	.51**

*$p < .05$ (two-tailed). **$p < .01$ (two-tailed).

What is the relationship among students' cognitive processes, online behaviors, and task performance? Table 6 presents correlations between participants' cognitive and online processes. The most striking finding of these analyses is the degree of internal consistency and the magnitude of the correlations between what participants were doing and their thinking processes. That is, success on the True Roots task, as measured by percent solved or the number of successful solve attempts, was correlated significantly with substantive cognitive processes (accurate cause-effect inferences, accurate evaluation of information). This finding is consistent with the idea that the IMMEX task evokes causal (i.e., scientific) thinking from learners. Poorer performance (but not necessarily lack of success) on True Roots, as measured by the number of unsuccessful solve attempts, was related significantly and negatively with substantive cognitive processing (accurate cause-effect inferences, accurate evaluation of information), and positively with confusion and inaccurate evaluation of information.

Table 6. Spearman Correlations Between Cognitive Processes and IMMEX Online Measures (n = 46)

Cognitive process[a]	% solved	Number of solve attempts: Successful	Number of solve attempts: Unsuccessful	Type of test[a]: Redundant	Type of test[a]: Lab/ data related	Type of test[a]: Experts	Type of test[a]: Library	Type of test[a]: Dictionary
1. Paraphrasing or echoing text	-.13	-.23	-.01	.29	-.44**	.46**	.43**	.21
2. Accurate cause-effect inferences	.52**	.68**	-.45**	-.32*	.58**	-.29	-.51**	-.37*
3. Inaccurate cause-effect inferences	-.10	-.05	.21	.08	.17	-.12	-.41**	-.31*
4. Accurate evaluation of information	.40**	.52**	-.52**	-.27	.42**	-.16	-.13	-.29*
5. Indeterminate evaluation of information	.36*	.34*	-.08	-.18	.34*	-.27	-.45**	-.24
6. Inaccurate evaluation of information	-.15	-.02	.48**	.02	.03	-.14	-.49**	-.09
7. Clarifying gaps in knowledge	-.23	-.37*	.09	.11	-.52**	.13	.68**	.60**
8. Judgments of information relevancy	-.17	-.16	.10	-.10	-.05	-.03	.04	.00
9. Confusion with content	-.23	-.32*	.32*	.28	-.19	-.14	.29	.35*
10. Awareness of failure to understand	-.20	-.20	.10	.23	-.18	.01	.23	.29
11. Verifying information	.02	-.01	.15	-.09	.01	.07	-.21	-.23
12. Monitoring problem solving behavior	.21	.25	-.03	-.06	.13	.03	-.25	-.19

[a]Proportion measure.
*p < .05 (two-tailed). **p < .01 (two-tailed).

A second set of relationships that offer strong support for the cognitive validity of True Roots is seen in the relationships between cognitive processing and the behavioral online measures. These relationships were consistent with prior studies[37]. Participants' use of data/lab tests was positively related to successful solve attempts, and accurate cause-effect inferences and evaluations of information, and negatively related to the use of paraphrasing or echoing text and clarifying gaps in knowledge. Interestingly, the use of the library and dictionary resources were negatively related to the use of scientific reasoning processes and positively related to the proportion of statements reflecting simple processing, confusion, or clarification of gaps in knowledge.

Summary. The set of cognitive processes identified as being evoked by True Roots provide compelling evidence for the efficacy of True Roots as a complex task demanding substantive cognitive processing from participants. Further, our findings that accurate cause-effect inferences and accurate evaluations of information (i.e., tests) were positively and significantly related to the number of successful solve attempts and negatively and significantly related to the number of unsuccessful solve attempts support the idea that scientific reasoning was an important factor in being successful on the task.

In addition, the set of relationships between cognitive processing and online process measures suggest a potential profile of participants. Successful participants engaged in proportionally more scientific reasoning than less successful participants. In general, successful participants relied on more laboratory/data tests and less on resource materials (i.e., experts, library, dictionary). Further, successful participants exhibited less confusion and had less of a need to fill gaps in their knowledge. In contrast, less successful participants misinterpreted information, were confused, and were unsuccessful at eliminating parents even though the tests they were using could have eliminated a parent.

DISCUSSION

In this study we gathered evidence of the types of cognitive processing participants evoked while engaged in the IMMEX True Roots task and examined the relationship among cognitive processes, online behaviors, and task performance.

Limitation of this study. The main limitation of this study is in the finding of a possible reactive effect of the think-aloud procedure. Participants' self-reports of the ease (M = 3.73, SD = 0.94, 1 = *not easy*, 5 = *very easy*) and completeness (M = 3.75, SD = 12.9, 1 = *not complete*, 5 = *very complete*) of the think-aloud procedure indicate a possible perceived effect. In addition, some differences were found between the control and think-aloud conditions, particularly in a gender by condition effect in the sequence of steps participants used in the task. Thus, caution is warranted when interpreting our findings—the findings may not generalize to a non-think-aloud condition.

Cognitive validation of IMMEX True Roots. In general, our findings provide strong support for the cognitive validity of the IMMEX True Roots task. We found strong evidence that the task evoked scientific reasoning from successful participants—that is, successful performance was associated with accurate causal inferences, accurate evaluation of information, and use of data/lab tests. Just as interesting are the relationships found with respect to less successful performance: Less successful performance was associated with inaccurate causal inferences, inaccurate evaluation of information, confusion, and expressing the need to clarify gaps in knowledge. Furthermore, these processes were strongly related to the use of resources—library, dictionary, or experts.

Towards a cognitive process validation methodology for online assessments. This study tested a method to validate assessments using think-aloud and clickstream data. As our findings suggest, participants' clickstream were strongly related to their cognitive processes. This is an important finding because to date, most online assessments rely only on aggregated clickstream data if used at all. In contrast, this study synchronized examinee's moment-to-moment online behavior with their moment-to-moment cognitive processing.

However, not all clickstream data may be equally useful. We think the structure of the IMMEX task and user interface have several important characteristics that enabled the clickstream data to be useful. First, the general architecture of the IMMEX task provided opportunities for participants to demonstrate understanding and opportunities *to demonstrate lack of understanding*. That is, accessing the data/lab test is a necessary condition for solving the problem, and material in the library and dictionary provide necessary background information

for someone unfamiliar with the content. Accessing ancillary material (e.g., experts) is unproductive. Systematic access of all three types of information over time yields different information about the examinee.

The second key feature is that the IMMEX architecture unambiguously captures intentional acts. That is, how to access the data/lab tests is visible and obvious to the examinee, and thus its use can be inferred as an intentional act. This is a critical feature because presumably, the act of clicking reflects the result of the examinee's reasoning and judgment. Finally, the grain-size of the information presented to the user is unambiguous. Each screen presents a single topic to the learner (vs. multiple topics); thus, there is little doubt about the content the examinee is viewing.

The important point is that judicious design of the user interface provides unique measurement opportunities, particularly in measuring intentional acts and the subsequent confidence in inferences drawn about processes underlying those acts. To the extent that the solution of the problem is tied directly to the information in the resource and less to prior knowledge or other individual difference variables, we think participants' ongoing use of the resources coupled with their ongoing measures of performance can be used as proxies for scientific reasoning, poor reasoning, and learning of content.

CONCLUSION

As the use of computer-based instruction increases, we expect assessment functions to be embedded into the instructional application. As engineering schools move toward ABET/EC2000 compliance, we anticipate a movement toward gathering a variety of evidence (vs. a single grade or survey), increasingly from on-line performance-oriented tasks, to better uncover what students are learning, the depth of their learning, and the process they are using to learn. One potential method to simultaneously satisfy instructional and assessment goals is to embed assessments within the instruction. Our findings demonstrate that students' online behavior are systematically related to their cognitions, offering a promising approach to measuring the cognitive processes underlying complex performance.

REFERENCES

[1] Accreditation Board for Engineering and Technology. (2000). *Criteria for accrediting engineering programs*. Baltimore, MD.

[2] American Society for Engineering Education. (1996). *Assessment white paper: A framework for the assessment of engineering education*.

[3] American Society for Engineering Education. (1998). *How do you measure success? Designing effective processes for assessing engineering education*.

[4] Waters, R., & McCracken, M. (1997). Assessment and evaluation in problem-based learning. In *Proceedings of the annual Frontiers in Engineering Education conference*, 689-693.

[5] Board on Engineering Education. (1995). *Engineering education: Designing an adaptive system.* Washington, DC: National Academy Press.

[6] Coward, H. R., Ailes, C. P., & Bardon, R. (2000). Progress of the engineering education coalitions. SRI

International, Arlington, VA, final report to the Engineering Education and Centers Division, NSF.

[7] Dowell, E., Baum, E., & McTague, J. (1994). *Engineering education for a changing world*. Washington, DC: American Society for Engineering Education.

[8] Meyers, C. W. (1995). *Restructuring engineering education: A focus on change*. National Science Foundation, Arlington, CA, report of an NSF Workshop on Engineering Education.

[9] Banios, E. W. (1991). Teaching engineering practices. In *Proceedings of the annual Frontiers in Engineering Education conference* (pp. 161-168).

[10] Wulf, W. A. (2000). How shall we satisfy the long-term educational needs of engineers? In *Proceedings of the IEEE*, 593-596.

[11] Chung, G. K. W. K., & Baker, E. L. (1997). *Year 1 Technology Studies: Implications for technology in assessment* (CSE Tech. Rep. No. 459). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

[12] Osmundson, E., Chung, G. K. W. K., Herl, H. E., & Klein, D. C. D. (1999). *Concept mapping in the classroom: A tool for examining the development of students' conceptual understandings*. (CSE Tech. Rep. No. 507). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

[13] Baker, E. L., Chung, G., Dennis, R., Herl, H. E., Klein, D., and Schacter, J. (1996). *Measurement of Learning Across Five Areas of Cognitive Competency: Design of an Integrated Simulation Approach to Measurement. Year One Report*. (CAETI Deliverable to ISX). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

[14] Klein, D. C. D., Yarnall, L., & Glaubke, C. (2001). Using technology to assess students' Web expertise. (CSE Tech. Rep. No. 544). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

[15] Schacter, J., Chung, G. K. W. K., & Dorr, A. (1998). Children's Internet searching on complex problems: Performance and Process Analyses. *Journal of the American Society for Information Science, 49*, 840–849.

[16] Bennett, R. E., Morley, M., Quardt, E., Rock, D. A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education, 13*, 303–322.

[17] Burstein, J. (2001, April). *Automated essay evaluation with natural language processing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

[18] Chung, G. K. W. K., O'Neil, H. F., Jr., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15*, 463–494.

[19] Chung, G. K. W. K., Harmon, T. C., & Baker, E. L. (2001). The impact of a simulation-based learning design project on student learning and teamwork skills. *IEEE Transactions on Education, 44*, 390–398.

[20] Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34*, 141–161.

[21] Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15*, 315–334.

[22] O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H.F. O'Neil, Jr. (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411–452). Mahwah, NJ: Erlbaum.

[23] O'Neil, H. F., Jr., Wang, S-L., Chung, G. K. W. K., & Herl, H. E. (2000). Assessment of teamwork skills using computer-based teamwork simulations. In H. F. O'Neil, Jr. & D. H. Andrews (Eds.), *Aircrew training and assessment* (pp. 245–276). Mahwah, NJ: Erlbaum.

[24] Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the cognitive science society* (pp. 412–417). Mawah, NJ: Erlbaum.

[25] American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

[26] Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice, 36*, 247-254.

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

[28] Chung, G. K. W. K., & Baker, E. L. (in press). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. E. Burstein, *Automated essay grading: A cross-disciplinary approach*. Mahwah, NJ: Erlbaum.

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

[30] Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

[31] 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.

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

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

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

[35] Klein, D. C. D., Chung, G. K. W. K., Osmundson, E., Herl, H. E., & O'Neil, H. F., Jr. (2001). *Examining the validity of knowledge mapping as a measure of elementary students' scientific understanding*. (Final deliverable to OERI). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

[36] 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.

[37] 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.

[38] Stevens, R., Casillas, A., & Vendlinski, T. (2001, April). *Artificial neural network-based performance assessments using simulations.* Paper presentation at the annual meeting of the NCME. Seattle, WA.

[39] Casillas, A. M., Clyman, S. G., Fan, Y. V., & Stevens, R. H. (2000). Exploring alternative models of complex patient management with artificial neural networks. *Advances in Health Sciences Education, 5*, 23–41.

[40] Kanowith-Klein, S., Burch, C., & Stevens, R. H. (1998). Sleuthing for science. *Journal of Staff Development, 19*(3), 48–53.

[41] Palacio-Cayetano, J., Allen, R. D., & Stevens, R. (1999). Computer-assisted evaluation–The next generation. *American Biology Teacher, 61*, 514–522.

[42] Palacio-Cayetano, J., Kanowith-Klein, S., & Stevens, R. (1999). UCLA's outreach program of science education in the Los Angeles Schools. *Academic Medicine, 74*(4), 44–47.

[43] Stevens, R. H., & Najafi, K. (1993). Artificial neural networks as adjuncts for assessing medical students' problem solving performances on computer-based simulations. *Computers and Biomedical Research, 26*, 172–187.

[44] Vendlinski, T., & Stevens, R. (2000). The use of artificial neural nets (ANN) to help evaluate student problem solving strategies. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences* (pp. 108–114). Mahwah, NJ: Erlbaum.

[45] Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations*. Los Angeles: University of California, National Center for Research on Evaluation Standards, and Student Testing.

[46] Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice-Hall.

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

[48] O'Neil, H.F., Jr. (1999). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior, 15*, 255–268.

[49] Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT.

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

[51] Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of Factor-Referenced Cognitive Tests*. Princeton, NJ: Educational Testing Service.

Acknowledgements

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, as well as by the Atlantic Philanthropies, Application Number 3977. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, the U.S. Department of Education, or the Atlantic Philanthropies. We wish to acknowledge and thank Joanne Michiuye of UCLA/CRESST for her help with the preparation for this manuscript, and Cecile Phan for her help with the scoring and transcribing of the audio tapes.

BIOGRAPHICAL INFORMATION

GREGORY K. W. K. CHUNG
Dr. Chung is a Senior Research Associate at CRESST. His work involves developing problem-solving assessments for computer-based delivery, and conducting research on the measurement of cognitive processes in online environments. Dr. Chung earned a Ph.D. in Educational Psychology from UCLA, 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.

LINDA F. DE VRIES

Linda F. de Vries is a Research Associate at CRESST. Her work at CRESST involves conducting research on computer-based assessments of problem solving and early literacy skills. Ms. de Vries earned an M.S. degree in Educational Science and Technology, University of Twente, The Netherlands.

ALICIA M. CHEAK
Alicia M. Cheak is a Research Associate at CRESST. Her work at CRESST involves developing computer-based assessments of children's early literacy and conducting research on the measurement of problem-solving processes. Ms. Cheak earned an M.S. degree in Psychological Studies in Education from UCLA, and a B.A. degree in English from UCLA.

RONALD H. STEVENS
Dr. Stevens is a professor of microbiology and immunology at UCLA and the developer and original programmer of the IMMEX problem-solving software environment. IMMEX is used in over 70 universities, medical schools, and K-12 schools. Dr. Stevens' research with IMMEX has focused on using IMMEX as a problem-solving assessment in multiple content areas using artificial neural network analyses.

WILLIAM L. BEWLEY
Dr. Bewley is the Assistant Director of Technology at CRESST. His research focuses on computer-based assessment, decision-support systems, information visualization, and distance learning. Dr. Bewley's background is in cognitive psychology and computer science, and he has managed numerous software development projects for military and educational training applications.