Practical Assessment in Complex Environments

Dr. Robert Pokorny, Dr. Jacqueline Haynes
Intelligent Automation, Inc.
Rockville, MD
bpokorny@i-a-i.com, jhaynes@i-a-i.com

Dr. Sherrie Gott
U.T. Health Science Center
San Antonio, TX
spgott@gmail.com

ABSTRACT

Assessment of complex performance in simulation environments is required for tailoring instruction, assigning competency levels or certifications based on performance in simulations, and evaluating training system effectiveness. Many systems require extensive research to create a scoring method for one scenario. Practical Assessment in Complex Environments (PACE) is a method that uses the collective wisdom of experts collected during reviews of trainees' performance to develop an objective scoring system that (a) correlates well with experts' holistic job assessment, (b) identifies performance weaknesses that guide future remediation, and (c) is easily administered, given a record of a trainee's performance in the simulation environment.

The scoring system requires that data be collected from trainees' performance within a simulation. Experts then review the samples, rank order the performances, and assign scores to reflect the quality of each sample. To capture the policies that each expert used, a panel of experts discusses the factors that led to each sample's score; typically, these factors are violations of good practice. A training psychologist links experts' critiques to elements of a prior cognitive task analysis, assigning point deductions to specific features consistent with the views expressed by experts.

When assessing the diagnostic skill of maintenance technicians troubleshooting faults deep within complex equipment simulation, PACE scores were valid: across various tests, they correlated in the .70s with time on the job. PACE scores were also reliable: after a scoring system for a scenario was created, when it was applied to new samples, the PACE scores correlated in the .80s with experts' holistic scores. PACE integrates the collective wisdom of experts within specific simulation contexts using a domain framework resulting from a cognitive task analysis.

ABOUT THE AUTHORS

Robert Pokorny has worked primarily in the areas of instructional environments and simulation since graduating with a Ph.D. in Experimental Psychology from the University of Oregon in 1985. At Intelligent Automation, he has worked on simulation-based visualization instruction, and directed projects in simulation-based battlespace information training systems and integrated content development and knowledge management. When working at the Air Force Research Laboratory, he focused on simulation-based Intelligent Tutoring Systems.

Jacqueline Haynes is co-founder, Executive Vice President and Director of the Education and Training Technology Group at Intelligent Automation, Inc. Her background combines education and psychology with AI applications. She received her Ph.D. from the University of Maryland in Curriculum and Instruction, and did post-doctoral work there in artificial intelligence and intelligent tutoring systems. Previously she was a faculty member at the University of Maryland, College of Education. Her research interests include research-based instructional design, tools for Web-based instruction, and reading comprehension.

Sherrie Gott is a cognitive/educational psychologist with over 25 years experienced in (a) analyzing expert and novice performance s in complex domains, (b) developing tools for assessing cognitive processes in such environments, and (c) developing tutoring systems based on principles of artificial intelligence to accelerate the acquisition of complex skills. She received her PhD in Educational Psychology form the University of Texas at Austin in 1979.

Practical Assessment in Complex Environments

Dr. Bob Pokorny, Dr. Jacqueline Haynes
Intelligent Automation, Inc.
Rockville, MD
bpokorny@i-a-i.com, jhaynes@i-a-i.com

Dr. Sherrie Gott
U.T. Health Science Center
San Antonio, TX
spgott@gmail.com

FUNCTION OF ASSESSMENT IN COMPLEX ENVIRONMENTS

Every simulation, game, or immersive environment is faced with similar, practical assessment questions:

- Does training improve performance? If so, how much?
- In what ways are trainees' performance improved by the training system?
- Can the trainees be certified to a standardized level of performance?
- How can assessment results be used to inform judgments of a trainee's competency?
- How should the training system guide further instruction?
    - Can the assessment guide effective instructional remediation and coaching during trainee's performance in the simulation?
    - Can the assessment guide instructional comments during the After Action Review?
    - Can the assessment inform future scenario selection or activities in upcoming simulation?

Difficulties faced by Assessment Systems

While assessing knowledge can use simple multiple choice or fill-in short answers tests, assessing the application of knowledge in performance settings should use performance tests. The fundamental difficulty in assessing an individual's performance is a many-to-one relationship between (1) performance and (2) the knowledge underlying that performance. That is, the same performance may result from people with different knowledge and characteristics. As a simple example of how performance may be due to many underlying causes, consider the case of a user of a new TV. The user may fail to setup the remote control correctly because (a) the user does not know that the remote control must be linked to the TV; (b) the user does not know that reading the manual will provide procedures for linking the remote to the TV; (c) the user reads the manual, but does not interpret the instructions to link the remote to the TV correctly; or (d) the user cannot read the manual. Correct diagnosis of the training need would lead to better training content and outcomes.

While acknowledging the probabilistic relationship between performance and cognitive processes, we sought to understand how domain experts assessed performance in complex environments: we began by collecting experts' reflections on others' performance records. Experts' assessed actions that may have many potential causes by interpreting actions in the best possible light; experts required clear evidence of unskilled performance before ascribing weaknesses to the performer. After capturing the policies that experts used to assess performance, we represented these policies in explicit scoring rules, and used them to score and analyze complex performance. This paper presents (1) how this approach was developed, (2) comparisons to other approaches, (3) results from applying the method, including validity and reliability measures, and (4) how this method can be applied to construct feedback and coaching.

This approach was begun to answer concrete questions about the effectiveness of an Intelligent Tutoring System (ITS) project sponsored by the Air Force Research Laboratory (See Gott and Lesgold, 2000, for a comprehensive review). The ITS taught diagnosis of faulty, complex equipment. This context led to specific requirements for the assessment system. First, the assessment system had to score performance samples of ill-structured problems (Simon, 1973). Dynamic, ill-structured problems involve multiple solution paths for complex diagnostic tasks. Experts varied widely in their approach based on their own individual knowledge and history; the assessment system allowed equivalent quality scores for many paths through the solution. Second, the assessment system should collect data in a context similar to the actual job, so results

could generalize from the test to job performance. Additional questions that could be asked of performers to elucidate their reasons for their actions were not asked because such questions might have led performers to think differently in the test environment than in the job environment, and the intent was to generalize from the test to on-the-job performance. Thus, the assessment system (a) allowed performers to take any action on the equipment they wanted to try at any time; and (b) performers were only asked to generate actions.

This paper provides an overview of three approaches to assessment: (1) PACE, the policy capturing we used, (2) Bayesian nets, and (3) linking job analysis to competencies. Then we will describe policy capturing with details and implications.

APPROACHES

Practical Assessment in Complex Environments (PACE) is a method of performance assessment in which experts review actions taken by performers in complex environments, and critique that performance. PACE collects critiques, standardizes them, and structures them within a goal structure of the overall task. PACE assigns point values (+ or -) to each critique. A student's performance on each scenario is calculated after calculating students' actions as fitting a critique, and then calculating an overall performance by summing point values from all critiques on that scenario.

Two other approaches used to assess performance in complex simulations are Bayes Net, and linking work analysis to essential competencies.

Bayes' Net approaches link actions taken by a student in a simulation to a set of underlying psychological constructs. Each construct has a score reflecting the best current estimate of the student's capabilities on that construct. A student's score on each construct changes as the student's actions fire rules that increase or decrease a student's score on the constructs. This approach requires specifying a set of constructs, and specifying rules that associate actions with constructs. The scoring system rules link actions to each construct by some probability. Determining these probabilities for all the rules is complicated. This Bayes' Net approach to scoring complex performance has been studied extensively in an effort to demonstrate that assessment of performance in simulations can have the same psychometric properties as tests used in simpler environments. This approach is called 'Evidence Centered Design' (Mislevy and Riconscente, 2005).

A second approach to measuring performance in complex simulations is work analysis and Mission Essential Competencies (Alliger, et. al., 2007). To define the training needs of warfighters in complex environments, the Air Force Distributed Mission Operations project has created an intensive set of methods and guidelines to relate job performance to categories of Mission Essential Competencies, Supporting Competencies, Knowledge & Skills, and experiences in which those other elements are trained. Scores of these constructs are based on performance, and then used to assign scores to the multi-tiered Mission Essential Competency model.

PACE differs from these other approaches in that PACE does not require the extensive research to identify underlying constructs, nor to associate the constructs with performance. Thus, PACE requires less work to develop assessments. Besides requiring fewer resources to develop an assessment system, PACE can construct feedback to users efficiently as well: it can use the link between students' actions and the experts' critiques to point out student flaws. Systems with more complex scoring systems also required more complex feedback systems. They would survey the psychological constructs, select one or more to discuss, and then construct the feedback text, linking the constructs to a student's actions. It should be clear that PACE is simpler than many alternative approaches. The rest of this paper will describe how PACE works in more detail, and report on how it addresses instructional and organizational goals.

PACE captures assessment policies of experts by the following approach: Performance samples are reviewed by experts, who score the samples based on overall quality judgments. Experts make their policies explicit during discussions of scoring with other experts. These policies identify violations of good practice. Violations are placed into a framework based on a cognitive task analysis of the domain. An explicit scoring system is created by identifying performance indicators that identify when a violation of good practice has taken place. Point deductions are assigned based on the severity of the violation. These point values are modified for each violation so that the overall score from the explicit scoring system is well correlated with experts' holistic judgments. This scoring approach results in reliable and valid scores, is easy to construct, and scoring policies are easily expressed.

PACE METHOD

PACE was constructed for a specific AFRL need: The data to which the expert policy capturing method was applied was very difficult troubleshooting of complex equipment. This equipment diagnosis which PACE evaluated and which the ITS taught involved maintenance activities in which units from a jet diagnosed by flightline maintenance were taken to a shop where they were tested for component fault isolation. In the shop, the unit is connected to a test station; the maintainer follows procedures to systematically test the unit under test (UUT) by injecting signals into the UUT, and measuring the unit's output. The maintainers' most difficult diagnosis occurs when the test station itself is faulty. The ITS taught and PACE evaluated maintainers isolating faults in the test station.

Each troubleshooting scenario for this training environment involved one fault; technician/trainees took many actions on the equipment to identify the problem. Actions yielded results in the testing situation just like they would on the real equipment. To test the Intelligent Tutoring System's instructional effectiveness, we compared the troubleshooting performance of technicians who received instruction using the ITS with technicians who had not. The performance data was analyzed by experienced technicians to make definitive and practical assertions about trainee performance, and the effectiveness of the training system.

The approach began by asking experts to make overall quality judgments of work performance records. Work performance records consisted of an initial symptom, followed by a set of troubleshooting actions and results that ended either when the fault was found, or the time limit (an hour) was met. Work performance records were collected, and formatted to be easy for experts to understand. (See Table 1)

Table 1. Example data from complex troubleshooting

Original Symptom: Fails test #21, 0 Volts DC	
ACTION	**RESULT**
Swap UUT and Rerun test	Test fails
Ohm check through active path of Test Package	Ohm tests all < 1 ohm
Replaced DMM fuses, rerun tests	Test fails
Voltage TestA1A3A12-11, 12	0 VDC
Voltage Test A1A3A12-47,48	5 VDC
Voltage Test A1A3A12-33,28	28 VDC
Swap A1A3A12 and rerun test	Test passes

Experts were then asked first to rank order the records on overall quality, and then assign quality scores, between 0 and 100. These experts' judgments were reliable and valid. To test reliability, multiple raters holistically scored sets of performance samples; the correlations between the raters were typically in the .80s (Pokorny, Hall, Gallaway, & Dibble, 1995). Thus raters could agree on quality of performance samples. To test validity, experts' holistic scores were correlated with time on job. Troubleshooting faults within the test equipment was a difficult task, and technicians with more than 10 years of experience were still improving. The performance scores from troubleshooting difficult faults correlated .75 with time on job (Gott, 1998). This correlation of time on job with experts' holistic scores indicates that holistic scores from performance records were a valid measure of job performance.

Apply Policy Capturing To Assess Performance As Experts Do

While it was important to determine that experts could reliably and validly score performance work records, a practical difficulty in using expert judgments is that experts might not be available when performance records need to be scored. This was the case with the Air Force ITS project, when the research to study the effectiveness of an ITS needed pre-test scores at the beginning of the study to check for similar scores between experimental and control groups. To apply expert judgment to assessing performance when experts were not available, the following two-step method of policy capturing was followed.

(1) After experts assigned quality scores to work performance records using the methods described earlier, we began to capture experts' policies by having them discuss the scores they assigned to performance records with each other. To inspire this conversation, psychologists selected the work performance records on which experts differed the most, and had them discuss why they scored the records as they did. Our experts, when asked to discuss the rationale behind their scores, normally expressed their reasons as critiques of performance.

(2) After capturing the rationale behind experts' scores, we created scoring worksheets that calculated scores given new work records. To create these scoring worksheets, we used a framework from a Cognitive Task Analysis conducted previously to inform and guide the construction of the ITS. The CTA identified mental models of equipment, and categories of procedures that were used to categorize actions for scoring. The structure used is shown in Table 2. The structural framework was based initially

on the structure of the equipment from experts' mental models: equipment was divided into the major equipment sections. One class of criticisms was incorrect jumps from one section to another. A second class of criticisms was investigating the wrong equipment within a section. The third class of criticisms was investigating correct equipment, but using incorrect procedures.

.

Table 2. Example Scoring Worksheet (partial)

Investigate Equipment: Unit Under Test	+25
Violations of transition between major units	
Skip to test equipment	-15
Premature jump to test equipment	-10
Violations of target in equipment in Unit Under Test	
Test off active path	-15
Test proving poor interpretation	-10
Procedural Violations within Unit Under Test	
Dangerous actions	-15
Uninformative actions	-10
Sub-optimal tests	-5
Investigate Equipment: Test Equipment—Signal Path	+45

After developing a worksheet, scores were calculated explicitly from the identified violations to mimic the scores assigned by human experts. Given that expert comments were primarily criticisms, most of the scoring rules identified point deductions. Points were granted to work performance records if they investigated equipment sections; other point rules were deductions associated with violations of good practice. The values of the deductions were initially based on the severity of the violation described by the experts in the discussion with other experts. After calculating overall scores from the point deductions, we compared the calculated overall score to human raters' overall scores. We adjusted the point values of the deductions so that the overall calculated value would be more similar to overall quality from human raters. This would continue until adjustments were not improving the overall correlation between calculated scores and human-rated scores.

To evaluate the adequacy of the captured scoring policy, we took new work performance records, and scored them using both the scoring worksheet and human holistic raters. Correlations between scoring worksheets and human raters ranged up to .89 (Pokorny, Hall, Gallaway, and Dibble, 1996).

Characteristics of the Scoring System

First, as mentioned above, the experts identified violations of good practices. Example critiques were items such as "the technician should not have started investigating the power supply yet," or "the technician is measuring wires between cards, which is inefficient."

Second, the critiques involved a sequence of actions that the experts believed definitively showed a trainee's misconception or misunderstanding. Thus, for many individual actions, expert reviewers reported that they would not change their assessment of a trainee's performance capability. Rather, expert raters considered sequences of actions and results, and would use the sequence to determine if a trainee's choices definitively indicated a misconception or performance weakness. Experts reported giving technicians the benefit of the doubt when it came to interpreting actions, only citing violations of good practice when they were clear. (This approach is at odds with many training systems' approach of updating a student model after each action.)

Third, the scoring worksheet was constructed by integrating the violations with the framework from the cognitive task analysis. For example, the task analysis identified equipment sections that were units, and types of procedures that could be attempted on each equipment section. The Cognitive Task Analysis results supplied a structure on which to attach specific violations of good practice.

RESULTS

This section describes the results of applying the PACE policy capturing method. When applied to ill-structured domain of complex system diagnosis, the results were valid and reliable. At the beginning of the article, we specified five practical questions regarding assessment to be addressed by PACE. Each of these is addressed in these results.

Question #1: Does the Training Improve Performance?

The first practical question in assessing the training system was "Did trainees' performance improve?" To answer this question, overall quality scores from trainees who received the ITS training intervention were compared with overall quality scores from trainees who did not. In the evaluation of the ITS for which this method was developed, all trainees participated in two work performance sample tests scored using the method we've described. The test results are shown in Table 3. Clearly the tutored students effectively benefited from the training intervention. The evaluation of this Intelligent Tutoring

System was described by Dr. Susan Chipman, recently retired Chief of Cognitive Science research at Office of Naval Research, as "one of the best evaluations of the effectiveness of intelligent tutoring that has yet been done." (Chipman, 2003). The comparison of overall quality of performance using the PACE scoring system was applied to a training study in this instance, but the PACE scoring system could compare performance differences between groups on any basis, not just training.

Table 3. Results from Troubleshooting Tests

Group	Pre-test		Post-test	
	Mean	S.D	Mean	S.D
Tutored	57	29	79	18
Un-tutored	53	22	59	20
Advanced technicians	82	12		

Question #2: What Elements Of Performance Were Improved By Training?

The second question identified in the Introduction was, "in what ways do performers improve as a result of the ITS?" With the availability of the explicit scoring policies that specify violations of good practice, training system administrators can easily compare post test performance of experimental and control subjects on each type of violation. Pokorny, Hall, Gallaway, and Dibble (1995) conducted this type of analysis. Explicitly capturing scoring policies enables an analysis of performance that identifies how two groups of performers, be they trainees or practitioners, differ.

One of the benefits of the explicit scoring policies was that facets of performance could be identified and evaluated for remediation by the training system. Thus, we identify which performance violations were improved by the training system, and by how much. Example results from this sort of analysis are shown in Table 4. Many other kinds of analyses, uses, and findings are possible from explicitly capturing experts' scoring policies.

Table 4. Comparing Tutored and Control Subjects on Facets of post-training performance

VIOLATION TYPES	% of scenarios in which violation occurred	
	TUTORED	CONTROL
Ohms inside Test Equipment	0	63
Jumping prematurely from component	27	56
Ohm tests dangerously	0	1
Targeting equipment that should be known to be good	3	42

DISCUSSION

PACE assesses complex cognitive performance in ill-structured domains. PACE scores correlate well with experts' holistic scores, which are valid and reliable. PACE can be used to measure the effectiveness of training systems (or to understand the difference between any groups of performers), and can be used to identify facets of performance on which those groups differ.

This paper began describing five practical questions. Two were addressed and reported in this paper. The remaining three practical questions were not directly addressed by the research. We will now discuss how policies captured from experts could be used to address certification, competency claims, and guiding future instruction.

Question #3: Can Captured Scoring Policies Be Used For Certification?

Frequently results of scoring systems are used to certify performers' capabilities. The policy capturing approach identifies violations of good practice within particular scenarios. It can be used to inform decisions to certify performance. Certifying performance capabilities requires (a) determining what quality of performance is required for certification, and (b) how do performance capabilities exhibited in the test generalize? Rules linking performance to certification could be designed as in Table 5. For each set of scenarios, the certification rules must link scenarios, and performance demonstrated in those scenarios, to certifiable performance capabilities.

Table 5. Linking Scenario Performance to Certification

Scenarios	Certification Levels
Scenario A1: investigates Section 1 Passing score 90%	Certification Level 1
Scenario A2: investigates Section 1 Passing score 90%	
Scenario A3: investigates Section 1 Passing score 90%	
Scenario B1: investigates	

Section 2 Passing score 90%	Certification Level 2
Scenario B2: investigates Section 2 Passing score 90%	
Scenario B3: investigates Section 1 Passing score 90%	

Question #4: Can Captured Scoring Policies Be Used For Determining Competencies?

As with certifications, performance of trainees assessed by policy capture can inform competency claims. Competency claims are similar to certifications, though they typically are seen as addressing more generalizable characteristics of the performers. Thus, competency will be considered attained when trainees demonstrate knowledge or behavior patterns across a variety of contexts. Performance evaluations are evidence of competencies, but competency claims will require rules that specify (a) how much evidence must be accrued from performance (either simulation-based or real world) and (b) contexts across which a behavior or knowledge is demonstrated before it is considered a true competency.

Question #5: Can Captured Scoring Policies Be Used For Guiding Future Instruction?

One of the most critical uses of assessments is in guiding future instruction. Describing how PACE can guide future instruction requires the most explanation of any of the five questions with which this paper began. PACE and the other two assessment approaches described earlier differ significantly in this area.

Guiding future instruction is divided into three activities: (1) before beginning a new simulated task, the system selects a new scenario or event which will help the trainee learn what he/she should learn next. (2) While performing within the simulation, the system coaches the trainee to complete the task. (3) After a task is completed, the system reviews the trainee's performance and guides trainees through a post-problem reflection. This is often referred to as an After Action Review. By computerizing the PACE scoring system, we can apply it to guiding future instruction. The approach of applying PACE to each of the three types of future instruction will be discussed in turn.

Future Problem Selection

Selecting future training scenarios, or events injected into an ongoing scenario can be informed by PACE. Future training should exercise the skills and knowledge identified as weaknesses in earlier training scenarios based on violations of good practice. To identify areas for future practice, the training system must have specified a learning trajectory which orders topics that trainees should learn (a simple learning trajectory would be addition, subtraction, multiplication, division). To identify content to train, future problem selection can use (1) the learning trajectory, and (2) observations of trainee performance which specify performance weaknesses. Additionally, future activities, events, or scenarios need to be indexed so that they can be selected when a violation of policies linked to a learning trajectory identifies a particular simulation experience which the trainee should encounter. See Table 6, which links violations of good troubleshooting practice to future scenarios.

Table 6. Linked violations of good practice to scenarios

Policy Captured: Violations of good practice	Scenarios or events that present opportunities
Needs to investigate Section 1	Put fault in Section 1
Needs to interpret procedures in tech manuals	Faults that require interpret procedures
Needs to interpret Voltage measurements	Faults require interpreting Voltage measurements

Coaching

Simulation-based training environments differ in how they coach. Coaching is the direction that an expert provides to a performer to improve the performer's performance. Three broad categories of coaching approaches are briefly described that differ in the depth of analysis of trainee errors.

The approach requiring the least analysis of student errors immediately directs the trainee to complete the next reasonable action; this coaching content will be given if the trainee asks for help, or if the training system deems the trainee requires immediate remediation. This coaching approach would be applied by a training system which embodies the philosophy of guiding trainees through the correct solution path (Koedinger and Anderson, 1993). This approach only requires that the system recognize deviations from an expert path; then any deviation from the specified path leads to a coaching response.

A second coaching approach provides trainees with a series of hints that ask the trainee to consider the

current situation in the simulation, and guides the trainee to take a reasonable action. This approach is used in training systems in which trainees can stray from optimal actions, and the training system provides assistance about how an expert thinks about the current situation (Gott and Lesgold, 2000).

A third approach, requiring the deepest analysis of student conceptions, provides coaching based on remediating psychological weaknesses that caused a trainee to make a particular mistake. Such an approach is used when the task performance is modeled in great detail, and the feedback to the trainee is based on a fine-grained analysis of student characteristics. Such an approach is used for systems such as Buggy (Brown and Burton, 1978).

The Policy Capturing approach is sufficient for guiding coaching with the first and second coaching approaches. The first coaching approach is based on what an expert would do at the current point, and does not adjust to what an individual trainee has done in the past. The second coaching approach gives feedback based on performance, and not on a deep modeling of the trainee capabilities. Hence, the policy capturing system of identifying violations of good practice would provide sufficient information for this coaching. The policy capturing approach would be insufficient for the third coaching approach, which requires modeling the underlying psychological constructs.

When ITS began, researchers believed that deeply modeling trainee psychological constructs could lead to providing just the right explanation that remediates individual student weakness (Wenger, 1987). However, contemporary instructional theory has led to a more sophisticated view of the instructional benefits of coaching since those early days of Intelligent Tutoring Systems. More recent studies of learning have moved from the trainee-as-consumer model of learning to trainee-environment—interaction model of learning. Sack, et. al, (1994), discussed their transformation from using student models in an attempt to transfer an expert's knowledge to students, to an understanding of students not as receivers of knowledge, but as constructors of knowledge. As an example of the type of instructional environment that contemporary instructional science is moving towards, Rosé, et. al. (2003) studied how asking "why" questions of students, which is known to produce important learning improvements, should be structured so as to increase the effectiveness of these constructive learning environments.

The implication of preferring the trainee-environment interaction model of learning compared to the trainee-as-consumer learning model means that, first, providing just the right content to a trainee is not as important as earlier thought. This finding is corroborated by studies of human tutors which show that tutors do not diagnose individual student weaknesses before providing tutorial intervention (Chi, et. al., 2003). Second, learning environments should provide interactions in which trainees reflect upon their knowledge in applied settings which leads to performance gains. Coaching systems could strive to provide educationally effective interactions rather than targeted explanations. Trainee-environment interactions do not depend on understanding each trainee's psychological deficiencies, but can be initiated by violations of good practice, as identified by PACE.

After Action Review

Constructing an After Action Review (AAR) involves two steps: (1) deciding what topics to discuss in the After Action Review and (2) developing content for system-student interaction on each topic. Deciding what topics to discuss is informed by the policy capturing approach; the policy capturing approach explicitly identifies action sequences that are violations of good practices with a specification of the severity of each violation. When reviewing the actions that a trainee has taken in the simulation, the violations are used to specify what topics should be raised in the AAR. Determining AAR content uses similar considerations for deciding what to present during coaching. Training interactions could follow one of the following approaches: (1) simply describe an expert's approach in the same situation in which the trainee performed sub-optimally; (2) provide questions and answers which help the trainee understand different choices and their bases within the simulation context; and (3) try to explain the trainee's underlying sub-optimal knowledge and skills, based on a deep diagnosis of the trainee's current cognitive structures.

Interactions with the trainee during AAR can be more extensive than interactions during coaching. In the middle of problem solving, trainees will be focused on achieving a solution. After the task is completed, though, the trainee can focus on one aspect of the just-completed scenario. Without concern for cognitive overload, the interaction can explore reasons that the trainee had for making particular choices, and elaborate more fully on sub-optimal choices.

Comparison of Assessment Approaches

Assessment by policy capturing (a) identifies the overall quality of a performance sample, (b) specifies

violations of good practice, and (c) can be used to guide future training.

The two other approaches described earlier, (1) a Bayes Net approach which links actions to underlying psychological constructs, and (2) links from work analysis to underlying competencies, share many characteristics, The Bayes Net approach sought to link actions to underlying psychological constructs. Evidence Centered Design, which evolved from a Bayes Net approach, aims to add psychometric rigor to those linkages of performance to underlying constructs. The approach linking work analysis to underlying competencies similarly link performance to underlying competencies. TBoth of these approaches require a great amount of labor to link actions to underlying constructs. These two approaches differ in their views of what the underlying constructs should be. The Bayes Net approach typically uses cognitive models which, if fully specified to support execution, could lead to the performance modeled. The work analysis approach specifies lists of competencies that enable performance, and lists of experiences which exercise those competencies.

PACE takes a more direct approach to linking actions to a scoring system. PACE takes critiques disclosed by experts, standardizes them into collections of critiques which are structured around the goal structure of the performers. This approach takes much less time to produce, and yields valid and reliable results. Further, when providing feedback to students and guiding future training interventions, the construction of these interventions is relatively easy, as they are based on actions students have taken.

Future Application of Practical Assessment in Complex Environments

Assessment by policy capturing identifies (a) the overall quality of a performance sample, (b) specific violations of good practice, and (c) based on explicitly linking performance deficiencies to experts' policies, performance assessment can easily inform instruction. We are planning to apply this practical policy capturing system to other domains, and to apply it to instructional systems in which it can be used to answer all five questions of practical assessment raised at the beginning of this report.

ACKNOWLEDGEMENTS

We thank our collaborators on the Air Force Research Laboratory Basic Job Skills project, including researchers from University of Pittsburgh, Dr. Alan Lesgold, Dr. Robert Glaser, and Dr. Sandy Katz; from McGill University, Dr. Suzanne Lajoie; from BBN, Dr. Allan Collins, Dr. John Fredericksen, Dr. Barbara White, and Mr. Bruce Roberts; from University of Michigan, Dr. Dave Kieras; from University of Colorado, Dr. Robert Linn; and from University of Massachusetts, Dr. Ron Hambleton. We thank the Air Force Leaders who had the foresight to support this project. Most of all, we thank the approximately 100 AF technicians who inspired us with their integrity, persistence, enthusiasm, and valor.

REFERENCES

Alliger, G., Beard, R., Bennett, W., Colegrove, C., and Garrity, M. (2007). Understanding Mission Essential Competencies as a Work Analysis Method. AFRL-HE-AZ-TR-2007-0034

Brown, J., and Burton, R. (1978). Diagnostic Models for Procedural Bugs in Mathematics. *Cognitive Science, 2*, p. 155—192

Chi, M.T.H., Siler, S.A. & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22, p.* 363-387.

Chipman, S. (2003) Overview: The U.S. Office of Naval Research Training Technology R&D. Presented at NATO Symposium. Downloaded from SITIS web site, http://www.dodsbir.net/sitis/ Dec 2005.

Koedinger, K. R., & Anderson, J. R. (1993). Reifying Implicit Planning in Geometry: Guidelines for Model-Based Intelligent Tutoring System Design. In S. P. Lajoie, Ed. & S. J. Derry, Ed (Eds.), *Computers as Cognitive Tools* (pp. 15-45). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Gott, S. (1998). Rediscovering Learning: Acquiring Expertise in Real World Problem Solving Tasks. Armstrong Laboratory/Human Resources Technical Report 1997-0009

Gott, S. P., & Lesgold, A. M. (2000). Competence in the Workplace: How Cognitive Performance Models and Situated Instruction Can Accelerate Skill Acquisition. In R. Glaser (Ed.), *Advances in instructional psychology*. Hillsdale, NJ: Erlbaum.

Hall. E.M., Gott, S.P., & Pokorny, R.A. (1995). A procedural guide to cognitive task analysis: The PARI methodology. Technical Report, (AL/HR-95-108). Armstrong Laboratory, Human Resources Directorate: Brooks AFB, TX.

Mislevy, R., Riconscente, M. (2005) Evidence-Centered Assessment Design. Principled Assessments for Design by Inquiry. Downloaded from http://padi.sri.com/downloads/TR9_ECD.pdf, Jun 22, 2010.

Pokorny, R., Hall, E., Gallaway, M., and Dibble, E. (1995). Analyzing Components of Work Samples to

Evaluate Performance. *Military Psychology. 8*, p 161-177.

Rosé, C.P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003). The Role of Why questions in effective human tutoring. In H. U. Hoppe, F. Verdejo and J. Kay (Eds.), *Artificial Intelligence in Education*. Amsterdam: IOS Press.

Simon, H. (1973). The Structure of Ill-Structured Problems. *Artificial Intelligence. 4* p. 181-201

Sack, W., Soloway, E., & Weingrad, P. (1994). Re-Writing Cartesian Student Models. In J. E. Greer & G. I. McCalla (Eds.), *Student Modeling: The Key to Individualized Knowledge-Based Instruction* (NATO ASI Series ed., Vol. 125, pp. 355-376). Berlin: Springer-Verlag.

Wenger, E. (1987). *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge*. Los Altos, CA: Morgan Kaufmann Publishers, Inc.