

# Developing and Implementing Computer Simulated Performance Tests

David J. Pucel, Ph.D.  
Professor Emeritus, University of Minnesota  
President, Performance Training Systems, Inc.

34<sup>th</sup> International Association for Educational Assessment (IAEA) Annual Conference 2008  
Cambridge University, UK. 9 September 2008

## Abstract

Testing performance capabilities to certify large numbers of people as competent to perform essential skills (e.g., medical, building) is costly and subject to questions of uniformity, validity, reliability, and quality control. With the advent of computer technology it is now possible to develop computer simulated performance tests in many areas. This presentation will present techniques for determining whether computer simulations are appropriate for testing selected skills, and the development and implementation of valid, reliable, and user friendly computer simulated performance tests. The techniques have been proven effective based on seven years experience with developing such tests, some of which are now used throughout the USA to certify personnel. The presentation will include the identification of test content, storyboarding the content for computer animation, development of the simulations with the aid of content experts, development of test tutorials to prepare people to take the tests, and gathering and evaluating test performance data. In each case theoretical, test development, and psychometric concerns will be presented along with examples of how they can be addressed.

## Introduction

Performance testing is the testing of a person's actual ability to perform a desired set of tasks or skills. It has long been an integral part of vocational and industrial training, and professional certification and licensing programs. Performance testing is a method of ensuring that people have the skills needed to perform competently.

Historically, performance testing was accomplished through apprenticeships and on-the-job training programs by a competent master training and observing a trainee as skills were being performed. Testing was informal and the assumption was that the master was competent to train and judge the competence of the trainee. The interaction between the master and the trainee continued for a long enough period of time for the master to train and judge the trainee as competent. Over time as the demand for larger numbers of skilled personnel increased, the apprenticeship system gave way to more formal training capable of training large numbers of people in the same skills at the same time. This required more formal performance testing methods. In addition, some occupations do not require the completion of formal training programs and accept people who have gained skills through work experience, self study, and other forms of informal preparation. In order to ensure competence those accepting these alternative methods of preparation have increasing been concerned about testing competence before allowing people to perform complex skills on their own. Governmental agencies have established licensure programs to ensure competence and protect the public, professional groups have organized certification programs to ensure professional competence, and businesses have developed testing programs to ensure competence of workers.

As the need for formal performance testing increased, performance test development and implementation procedures evolved. Pucel (2005, pp. 156-165) presents step-by-step procedures for the development of such tests.

This paper focuses on unique issues and procedures for developing computer simulated performance tests. Although there is substantial overlap between the development of hands-on performance tests and computer simulated tests, computer simulated tests present many additional issues and challenges. This paper includes the following:

- Identifying skills to be tested.
- Determining the reasonableness of testing through computer simulation.

- Developing the underlying performance test instrument.
- Establishing computer simulation design parameters.
- Developing the multi-media presentation.
- Preparing test takers.
- Validating the test.
- Data based analysis and refinement of the test.
  - Advanced validation
  - Establishing reliability

### Identifying skills to be tested

The skills to be performance tested are typically identified through a job analysis which identifies skills proven to be important to job performance. If the tests are to be used for employee selection or other forms of differentiating employees which affect their employment, this process must meet certain requirements within the United States as specified in the EEOC Guidelines (Uniform Guidelines on Employee Selection Procedures, 1978). For example, the group from which job analysis data are gathered must include job incumbents and other subject matter experts and must have gender, ethnic and racial group representation. Table 1 presents a portion of a job analysis for ophthalmic technicians. It presents a list of skills and their relative importance ratings as judged by subject matter experts. The analysis shows that “test for versions” is the most important skill. Therefore, it is an appropriate skill to be tested.

Table 1  
Job Analysis

Skill List	Importance Rating	Ranking
Test for versions	4.80	1
Determine the correction in patient lenses	4.76	2
Test for far vision	4.64	3
Measure pressure in the eye	4.17	4
Obtain informed consent	3.94	5
Measure pupil size	3.92	6

“Test for versions” involves testing to see if a person’s eyes track together. It is performed by asking a person to look at an object which is moved around the eyes to determine if both eyes track together.

### Determining the reasonableness of testing through computer simulation

Not all performance skills can reasonably be tested through computer simulated tests. Where hands-on performance testing allows evaluators to observe people completing all aspects of the performance, computer simulations are limited in what can be tested. For example, because people completing computer simulated tests typically manipulate equipment with a computer mouse, it is not possible to observe a person’s ability to physically manipulate the real equipment. The following are some guidelines for determining if it is reasonable to test a specific skill with a computer simulated test.

Computer simulations are a viable method of testing a skill:

- If people already have the necessary tactile or psychomotor skills (e.g. computer skills, real equipment manipulation skills).
- If the outcome desired is assuring a person can complete a process correctly.
- If the results can be recorded and evaluated within a computer. (Pucel & Anderson, 2006)

Computer simulations are typically not viable:

- If the actual physical handling of things associated with the skill is to be evaluated. (e.g., tactile touch of a camera)
- If the desired outcome is a physical product such as a cake that requires modifying materials.

- If the goal is to evaluate results yet the results can not be recorded in a form that can reliably be evaluated by a computer (e.g., a dance). (Pucel & Anderson, 2006)

In addition to determining the viability of testing a skill using computer simulations one must also determine if it makes sense from a business perspective. Some of the business advantages are:

- Standardization of testing & evaluation.
- Greater test reliability.
- Elimination of the need to train evaluators & test subjects.
- Reduction of evaluation bias.
- Long-term cost savings (both to those administering the tests and people taking the test, e.g., travel time).
- Reduction of potential risk of litigation. (Pucel & Anderson, 2006)

Some of the disadvantages of computer simulated performance testing are:

- Substantial developmental costs.
- Lengthy test modification time.
- Complex validation processes.

### Developing the underlying performance test instrument (checklist)

The development of any formal performance test, whether it is a hands-on or computer simulated test, begins with clarifying the skill to be tested, and how its performance will be judged. The process typically starts with the development of a performance checklist which is the underlying test instrument (Pucel, 2005, pp. 156-165). Table 2 presents a performance checklist for the skill: test for versions.

Table 2  
Test for Versions Performance Checklist

Conditions: A patient with vision in both eyes

Skill: Test for versions

Standard: Minimum score of 50 out of 75

Actions	Criteria	Satisfactory	Unsatisfactory
1. Seat patient.	Patient seated in examination chair	5	0
2. Position self	About 15 inches in front of patient	5	0
3. Instruct patient	Follow my finger with both eyes	15	0
4. Move finger	Through six cardinal positions of gaze	25	0
5. Record results.	Normal – Abnormal ____ If abnormal Which eye Right ____ Left ____ Muscle effected _____	25	0

(Pucel, 2008)

Based on Cassin & Hamed 1995

The checklist details:

1. The conditions under which the test is to be conducted.
2. The process actions involved with performing the skill.
3. The criteria which indicate how the performance of each action will be judged.
4. A column indicating the points that will be awarded for performing each action satisfactorily.
5. An unsatisfactory column for recording unsatisfactory performance of each action.
6. The cut score that will be used to judge pass or fail.

At times the speed or timing of the performance is also scored. In the example presented in Table 2 time is not important.

If the performance test was to be used in a hands-on testing situation next steps would be to plan how the test would be administered.

- Whether the test is to be given within an actual job environment or a simulated environment such as a classroom.
- The qualifications of observer evaluators.
- The number of evaluators that would evaluate each person.

### **Establishing computer simulation design parameters**

There are many more design considerations upon deciding to use computer-based simulated performance tests than when using traditional hands-on performance tests. Typically no one person has the expertise to adequately address the entire range of design considerations. Therefore, a developmental team needs to be assembled. The team includes a range of expertise represented by:

- Test development organization management.
- Subject matter experts.
- A test designer.
- Multi-media developers.
- A psychometrician.

The design considerations can be divided into macro considerations which affect the overall simulation and micro design considerations which affect how the macro considerations will be implemented. Some of the macro design considerations are as follows.

- How much fidelity/realism is necessary?
- What scenario(s) and features of the real-world environment will be simulated?
- How will the examinee interact with the simulation? How will examinees be guided through the simulated scenario?
- Will the simulation have a static path or be open/dynamic?
- What will the computer record as potentially scorable information?
- How will the scorable information be combined or will the score simply be based on the final outcome? (process vs. product/result)

(Knapp & Pucel, 2008)

In addition multi-media software must be selected based on its flexibility and availability at test sites.

### **Developing the multi-media presentation**

Micro design considerations continually present themselves during the actual development of the simulated tests. They affect what is actually tested and how it is tested. For example, will examinees be asked to enter results information in open ended blanks or will they be asked to select from result choices? Will the equipment be fully functional with all components being manipulable throughout the simulation or will only reasonable options be allowed? The answers to each of these questions affect the costs of creating the simulation but also they affect the depth of testing. If result choices are presented, the examinee is more likely to guess the correct answer. If all components are not manipulable, the possibilities for performing incorrectly are limited.

The process of developing multi-media simulations begins with the development of a storyboard which includes the content of the simulation and how that content will be presented. The storyboard follows the underlying performance checklist developed earlier. Storyboarding a computer simulated performance test is much more involved than storyboarding an instructional program. In addition to presenting the correct way of doing things one must also present plausible incorrect ways. This is similar to presenting the correct answer and incorrect answers in a multiple choice test. In addition, the storyboard must include a description of the visual images to be presented. Table 3 presents a partial storyboard for the skill "test for versions." Notice it presents the performance actions, the correct response to each action, plausible incorrect responses to each action, and what should be included in the visual image of each action. The storyboard provides the multi-media developer with a set of developmental specifications.

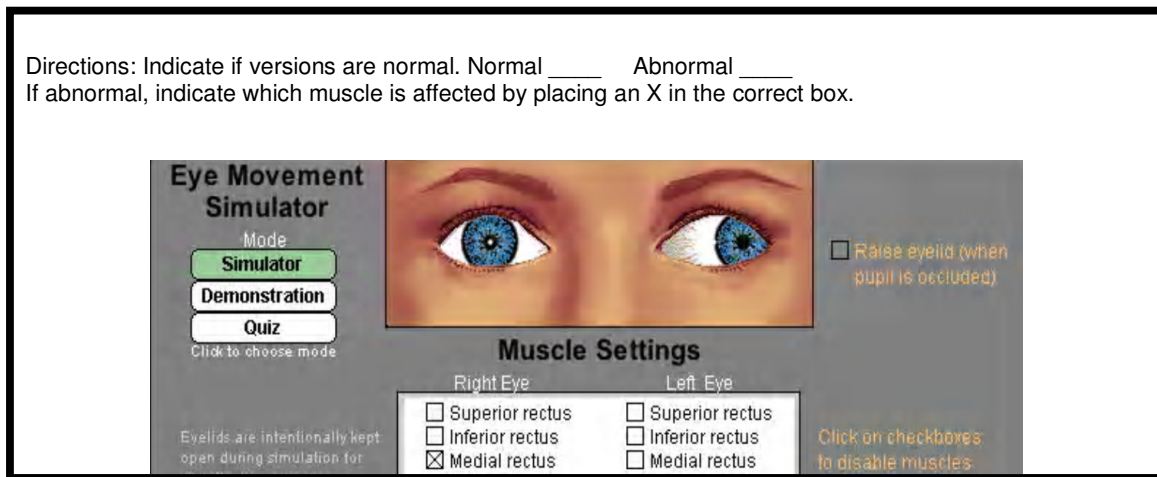
Table 3  
Test for Versions Storyboard

Incorrect	Actions	Correct	Visual Image
Patient not directed to be seated	1. Seat patient	Patient seated in examination chair	Patient's face and eyes appear on the screen
Examiner either too close or too far	2. Position self	About 15 inches in front of patient	Too close – eyes do not move. Too far- eyes do not move in the full range.
Patient does not follow finger	3. Instruct patient	Follow my finger with both eyes	Patient follows finger
One or more of the positions left out	4. Move finger	Through six cardinal positions of gaze	Right, upper right, lower right, left, upper left, lower left
	5. Record results	Normal ___ Abnormal ___ If abnormal: Which eye? Right ___ Left ___ Muscle affected _____	

The multi-media developer usually begins the development by taking actual pictures and movies of people performing the skill according to the storyboard to obtain the visual images. Those pictures and movies are incorporated into a computer manipulated simulation using programs such as FLASH. Options to perform each action correctly and incorrectly are included. Also included is a scoring algorithm that incorporates the scoring based on the performance checklist.

Figure 1 presents a portion of the computer simulated performance test for the skill “test for versions.” It presents the simulated activity for the actions 4 and 5: “Move finger” and “Record results.” During the

Figure 1  
Sample Computer Screen of a Simulated Test for Versions



Adapted from the simulation developed by Rick Lasslo, M.D., M.S., Gary Henderson, Ph.D., and John Keltner, M.D. University of California, Davis, School of Medicine  
<http://cim.ucdavis.edu/eyes/version15/eyesim.htm>

actual simulation the person being tested moves the computer mouse through the six cardinal positions gaze which are right, upper right, lower right, left, upper left, and lower left. During the process the person

observes the eyes and if the eyes do not track together records which muscle is abnormal by checking the name of the muscle.

### **Preparing test takers**

People who take computer simulated performance tests are required to use substantially different psychomotor and tactile skills than are required to perform the actual skills in the real world. Typically they are expected to move objects with a computer mouse by placing the cursor on arrows, clicking on a joystick, or making choices by pressing certain keys. They are expected to record their results by making selections or by entering data with a computer keyboard rather than just writing them down on a piece of paper.

In order for the process of completing the test to not contaminate people's ability to demonstrate their true skill, people need to be shown and be able to practice the types of actions that will be expected during the test. Therefore, most computer simulated performance tests require the development of an extensive tutorial which is sent or presented to people prior to taking the test. The challenge is to provide the people with an opportunity to see the types of things they will see on the test and be able to practice moving things and entering data without giving them key information which the test is designed to test. In other words, people need to be allowed to familiarize themselves with the methods of taking the test and not how to perform the skill to be tested.

### **Validating the test**

When using hands-on performance tests validation is a relatively simple process because the actual performance of a person can be observed by content expert evaluators and it becomes quite apparent if multiple experts consistently report the test is not yielding results consistent with expert judgements. Validating computer simulated tests is much more complex. Often comprehensive data and information need to be assembled to provide evidence that the test is accurately assessing a person's true ability to perform. Without such evidence the profession and those being tested often associate not passing the test with flaws in the test rather than inadequate performance.

Validating a computer simulated performance test requires multiple levels of validation. They include:

- Content validity - The extent to which the test assesses the content to be tested. This is typically done by subject matter experts who were members of the developmental team. Although content validation is addressed during the development of the performance checklist, once the simulated test is developed subject matter experts must also address the content validity of the final test. Often this is accomplished by asking subject matter experts who were members of the developmental team to complete the test and judge the extent to which the actual simulated test does assess what it was intended to test.
- User validity - The extent to which users (examinees) perceive the test as representing the real world skill and its ability to measure their true skill. This is typically assessed through pilot testing with both people who already have the skill and people like those that will eventually be tested. Sample questions include:
  1. How realistic do you think what appeared on the screen represented the real world?
  2. Was what you saw on the screen similar to what you would see in actual practice?
  3. Could you get around easily within the simulation?
  4. Were you able to reverse mistakes?
  5. Do you believe the simulation test allowed you to demonstrate your true skill? (Pucel & Anderson, 2006)

Establishing user validity also applies to the tutorial as well as the performance test itself. During pilot testing people need to be asked about the extent to which the pre-test tutorial prepared them for what they actually experienced during testing.

- Scoring validity - The extent to which the test is scoring consistent with the underlying scoring algorithm as specified in the performance checklist. Establishing scoring validity is complex because there are potentially many places along the developmental process where scoring errors may be introduced. At times multi-media programmers do not understand the importance of scoring exactly how the test designer and psychometrician intend scoring to take place. Therefore, they take short cuts to make implementation of scoring easier. This can lead to scores not representing exactly what

they were intended to represent. Also, the multi-media programmers usually develop the system for communicating final examinee test results in the form of a data file. That file can be structured in many ways and at times it may not be structured to communicate all of the necessary information. An efficient way of establishing scoring validity is to pilot test with subject matter experts and have them review their scores on each test component against the test results generated by the simulated test. Consistent discrepancies can reveal scoring problems.

#### Data based analyses and refinement of the test

- Advanced validation
- Establishing reliability

There is increasing pressure to provide statistical data to support the adequacy of performance tests which are similar to statistics provided for multiple-choice tests. Based on this need to provide statistical data to support the validity and reliability of simulated performance tests, Pucel developed the Performance Test Analysis and Refinement System (PTARS) (© Pucel, 2007). Pucel's PTARS gathers results data on each scored component of a test and the overall performance of the test and judges the data against a set of established criteria and results obtained from independent skill competent evaluators. Data are used to analyze each test component as a basis for determining test validity, test performance, test refinement, and reliability estimation.

Pucel's PTARS (© Pucel, 2007) has three levels of sub-analyses.

Level 1: Analyze overall pass/fail rates (include all that apply)

- Overall combined process and results pass/fail rate.
- Process pass/fail rate.
- Results (product) pass/fail rate.

Level 2: Analyze individual process actions and results components pass/fail rates. (Similar to item analysis)

Level 3: Statistically compare computer scoring results with those from content expert evaluators to establish validity and estimate inter-rater reliability.

This system will be demonstrated around a familiar skill, "change a tire." The sample checklist for the skill is presented in Table 4. It was developed using the procedures described earlier. In this case the performance is to be judged both on the process of performing as well as the results of performing.

Table 4  
Sample Checklist for the Skill Change a Tire

Conditions: A car with a flat tire, tools, and a replacement tire

Skill: Change a tire.

<b>Process</b>	<b>Criteria</b>	<b>Satisfactory</b>	<b>Unsatisfactory</b>
1. Block the tires	Blocks in front and back of tires not to be raised	2	0
2. Loosen lug nuts	Loosen ¼ turn	1	0
3. Position jack	Under jack-point as in manual	1	0
4. Jack car	Raise until tire clears the ground	1	0
5. Etc.			
<b>Result</b>	<b>Criteria</b>	<b>Satisfactory</b>	<b>Unsatisfactory</b>
1. Replacement tire mounted	Tire that is on the car is not the one that was to be replaced.	1	0
2. Lugs tight	Reading on the torque wrench during the process of tightening the lug nuts is consistent with car manual.	1	0
3. Tools & tire replaced	Tools are returned to where they were before the test.	1	0

Cut score: 11

\* Critical Step

(Pucel, 2005)

PTARS Level 1 addresses the actual pass/fail rates associated with people taking the test. At times both the process of completing the skill as well as the final results from applying the process are judged. At other times only the process or the final results is judged. If both are tested, each is judged as well as the overall combination. One standard for judging pass/fail rates is to assume that if people come to testing with stated prerequisites, then 70% or more should pass the test. Although this standard is not a formal professional standard, it is often used with multiple choice tests and is used with this system.

Table 5 presents a sample analysis of the overall pass/fail data for people who completed the simulated performance test of the skill “change a tire.” This sample analysis reveals that the test performed well on both the testing of process actions and the results components, as well as overall using the criterion of 70% pass.

Table 5  
Sample Level 1 Analysis of Pass Data

Computer Decision	Total
Failed Overall	22
Failed Process	4
Failed Results	20
Pass Overall	120
Percent passed process	97.18%
Percent passed results	85.91%
Overall passed	84.51%
Total N	142

PTARS Level 2 is designed to perform the same function as an item analysis for a multiple choice test. It addresses the question: How well is each scorable component of the test performing? This is accomplished by analyzing the pass results for each of the individual process actions and results components. In the example presented in Table 6, even though the overall process pass rate was high as indicated in Table 5, there appears to be a problem with the process step “loosen lug nuts.” Upon further review it might be determined that people could not clearly see that process during the test and the simulation would have to be modified. Again, pass/fail data for each process action component and results component are judged against a standard of 70%.

Table 6  
Sample PTARS Level 2 Analysis for Change a Tire Simulated Test

Process N = 142	Block tires	Loosen lug nuts	Position jack	Etc	Overall process pass
N-correct	139	90	142		138
% Correct	97.89%	63.38%	100.00%		97.18%
Results N = 142	Replacement tire mounted		Tools & tire replaced	Overall results pass	
N-correct	115		127	122	
% Correct	80.98%		88.43%	85.91%	

PTARS Level 3 is the most advanced format for analyzing the performance of a simulated performance test. It is used to compare computer scoring results with those of content expert evaluators to review the concurrent validity of each scored process action and results component. It is also used to estimate inter-rater reliability by comparing the overall pass scores produced by the computer simulation with the overall pass scores for the same examinees as judged by content expert evaluators. Level 3 is conducted in three steps.

First, gather scoring data from the computer simulation and content expert evaluators on the performance of the same group of examinees. This is typically done as follows.

1. Have examinees complete the computer simulation test and record the real time computer screen images on video tape.
2. Record the computer scoring of each examinee's performance.
3. Have expert evaluators (i.e., N = 5) review the tape recorded information and score the performance of the examinees using the same scoring protocol (checklist) as was to be used by the computer.
4. Record the majority scoring decisions across the expert evaluators.

Second, compare the computer pass scoring results with those of the majority of expert evaluators to judge the concurrent validity of each scored component. Table 7 presents a table of such results. The table can be reviewed for practical significance by examining the magnitude of the actual differences, and statistically by using statistics such as Chi-square.

Table 7  
Sample PTARS Level 3 Comparison Computer and Expert Evaluator Component Pass Scoring (N=50)

<b>Scored Component</b>	<b>Computer</b>	<b>Majority SMEs</b>
Pass overall	40	37
Pass results	43	42
Pass process	45	40
<b>Process</b>		
Block tires	46	44
Loosen lug nuts	49	49
Position jack	50	50
<b>Results</b>		
Replacement tire mounted	50	50
Lugs tight	42	45
Tools & tire replaced	45	44

Third, compare the overall test pass results from the computer simulation with those of the majority of content expert evaluators to determine the inter-rater reliability of the test. Inter-rater reliability is defined as the extent to which two or more individual raters agree. This represents the consistency of the implementation of a rating system. In this case the two raters are the computer and the majority of the content expert evaluators. Table 8 presents a table of such results.

Table 8  
Sample PTARS Level 3 Inter-Rater Reliability Based on Overall Test Pass (N=50)

<b>Examine</b>	<b>Computer</b>	<b>Majority SMEs</b>
1	pass	pass
2	pass	fail
.		
49	pass	pass
50	pass	pass

The actual inter-rater reliability can be calculated by correlating the computer scores with the content expert scores using a statistic such as a Phi-coefficient.

### Summary

Performance testing of important skills is becoming more and more a necessity to ensure people can perform competently. The use of computer simulated performance tests is a viable way of testing large numbers of people at very diverse geographical locations. However, care must be taken to ensure that the skills can reasonably be tested with computer simulations and that there is a sufficient business case. The development of such tests requires substantial commitments of time and resources. Their development requires a wide range of expertise and extensive attention paid to establishing validity and reliability issues to ensure psychometric soundness. However, once they are adequately developed they become more cost effective the more they are used and they are less subject to testing errors due to such influences as evaluator bias, equipment variations, and unequal testing conditions which can make performance tests subject to legal challenges. Current simulated performance test development procedures and associated psychometric procedures have evolved and are available and are relatively easy to use although complex. A number of organizations have developed test development standards which are currently being revised to include more guidance on the development of performance tests (AERA, 1999) (Browning and Mullins, 1996).

### Bibliography

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington D.C.: American Educational Research Association, 1999. ISBN 0-935302-25-5
- Anderson, L.D., & Pucel, D.J. (February, 2007). *Computer simulated performance testing*. Annual meeting of the Association of Test Publishers. Palm Springs, CA.
- Anderson, L.D., & Pucel, D.J. (January, 2005). *Performance testing: Issues, methods and implementation*. Meeting of the Performance Testing Council. Reno, NV.
- Browning, A.H., Bugbee, A.C., & Mullins, M. (Eds.) (1996). *Certification: A NOCA Handbook*. Washington, DC: National Organization for Competency Assurance. [Look for 2<sup>nd</sup> edition later in 2008]
- Cassin, B., & Hamed, L.M. (1995), *Fundamentals for ophthalmic technical personnel*, ISBN-13: 9780721649313, Elsevier Health Sciences, Philadelphia, PA, 473pp.
- Knapp, D.J. & Pucel, D.J., (April 9, 2008) "Performance Testing: A New Frontier for IO Psychologists", Society for Industrial and Organizational Psychologists Annual Meeting Workshop, San Francisco, CA, 2008.
- Knapp, D.J. (2006). The U.S. Joint-Service Job Performance Measurement project: An enduring legacy in performance measurement. In W. Bennett, C.E. Lance, & D.J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Lasslo, R., M.D., Henderson, G. & Keltner, J. *Eye Simulator*, University of California, Davis, School of Medicine, <http://cim.ucdavis.edu/eyes/version15/eyesim.htm>
- Pucel, D.J. (2005). *Developing and Evaluating Performance-Based Instruction (third edition)*. ISBN 0-943919-03-7 New Brighton, MN: Performance Training Systems, Inc.
- Pucel, D.J., & Anderson, L.D. (October, 2006). *Computer simulation performance testing*. Proceedings of the Ninth IASTED International Conference on Computers and Advanced Technology in Education 2006 (CATE), ISBN Hardcopy: 0-88986-626-0 / CD: 0-88986-628-7, Lima, Peru., pp. 339-344.
- Pucel, D.J., & Anderson, L.D. (June, 2003). *Developing computer simulation performance tests: challenges and criteria*. Proceedings of the IASTED International Conference on Computers and Advanced Technology in Education 2003 (CATE), ISBN: 0-88986-361-X: Rhodes, Greece pp. 170-174.
- Uniform Guidelines on Employee Selection Procedures (1978). Section 60-3, U.G.E.S.P. (1978); 43 FR 38295 (August 25, 1978). (<http://www.uniformguidelines.com>)
- Yang, Y., Buckendahl, C.W., Juskiewicz, P.J., & Bholra, D.S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, 15, 391-412.