# Performance Assessment: Something Old, Something New

24

Dee H. Andrews and Wallace H. Wulfeck II

**Abstract**

Education, industry, and the military rely on relevant and effective performance assessment to make key decisions. Qualification, promotion, advancement, hiring, firing, and training decisions are just some of the functions that rely on good performance assessment. This chapter explores issues, tools, and techniques for ensuring that performance assessment meets the goals of decision makers. Important considerations include the following: the task environment in which assessment takes place, the validity of the measures selected, the diversity and scope of the measures, and often the local political environment in which assessments are made. Unfortunately, primarily in education, assessment policy is a matter of intense political debate, and the debate is sustained by misinformation. For example, traditional paper and pencil assessment techniques have come under fire from those who do not feel they have the relevance they once did. Simulation-based training technologies in industry and the military have put more emphasis on performance assessment systems that show real-world work relevance. The chapter examines these topics.

**Keywords**

Authentic assessment • Alternative assessment • Portfolio assessment • Performance appraisal • Performance evaluation • Performance task

## Introduction

In education, performance assessment refers to testing methods that require students to create an answer or product, or execute a process, that demonstrates their knowledge or skills. Performance assessment, in education and work settings, can take many different forms including writing short answers, doing mathematical computations, writing an extended essay, conducting an experiment, presenting an oral argument, executing a series of tasks, or assembling a portfolio of representative work (US Congress, Office of Technology Assessment, 1992). More broadly, performance assessment refers to the measurement of a system or process with respect to goals or benchmarks set for it.

Assessment in one form or another has a history going back at least 2000 years. In early China, prospective civil servants were examined not only on recitation, but on productive originality (Madaus & O'Dwyer, 1999). Later, the focus turned away from production toward reproductive thinking "because government officials became worried that the scoring of these questions would be too subjective; thus, they reverted back to questions that required more rote answers" (Madaus & O'Dwyer, 1999). Twelve hundred years later, universities in France and Italy began the practice of

D.H. Andrews (✉)
Arizona State University, Mesa, AZ, USA

425 E. Melody Lane, Gilbert, AZ 85234, USA
e-mail: dee.andrews@asu.edu

W.H. Wulfeck II
Space and Naval Warfare Systems Center Pacific, 53560 Hull St, San Diego, CA 92152, USA

12517 Fairbrook Rd, San Diego, CA 92131-2234, USA
e-mail: wally.wulfeck@gmail.com

J.M. Spector et al. (eds.), *Handbook of Research on Educational Communications and Technology*,
DOI 10.1007/978-1-4614-3185-5_24, 

oral examinations, and written examinations in Latin composition appeared by the sixteenth century (Madaus & O'Dwyer, 1999). Guild membership led to professional certification based on proven skills and knowledge. The industrial revolution brought a focus on quantification in measurement, and by the twentieth century industrial task analysis and "scientific management" had given rise to the invention of multiple-choice testing by Frederick Kelly in 1914 and its large-scale use in the Army Alpha test in World War I. The College Board's Scholastic Aptitude Test used multiple-choice items beginning in 1926. By the 1950s, the invention of data processing systems and optical scanners, as well as taxonomies of educational outcomes (Bloom, 1956) and the behavioral objectives movement, led to the predominance of machine-scorable tests, particularly in the USA.

Since at least the 1950s, there has been a continuing societal debate about the use, fairness, and appropriateness of various forms of assessment. Robert Glaser (1963) focused on the *purpose* of testing and distinguished between norm-referenced and criterion-referenced testing. Performance assessments are usually criterion-referenced because they refer to defined performances, but can be norm-referenced, for example in sales or sports. More recently there has been ongoing controversy concerning the proper role of testing in schooling, and this has tracked larger societal debates concerning the "constructivist" and "situated learning" movements (Anderson, Reder, & Simon, 2000), and most recently, standards-based education policy with so-called "high-stakes" testing.

## Definitions

*Authentic Assessment*: Engaging and worthy problems or questions of importance, in which students must use knowledge to fashion performances effectively and creatively. The tasks are either replicas of or analogous to the kinds of problems faced by adult citizens and consumers or professionals in the field. (Wiggins, 1993, p. 229).

*Alternative Assessment*: The term alternative assessment is broadly defined as any assessment method that is an alternative to traditional paper-and-pencil tests. Alternative assessment requires students to demonstrate the skills and knowledge that are difficult to assess using a timed multiple-choice or true-false test. It seeks to reveal students' critical-thinking and evaluation skills by asking students to complete open-ended tasks that often take more than one class period to complete. While fact-based knowledge is still a component of the learning that is assessed, its measurement is not the sole purpose of the assessment.

Alternative assessment is almost always teacher-created (rather than created by other test developers) and is inextricably tied to the curriculum studied in class. The form of assessment is usually customized to the students and to the subject matter itself. (Teaching Today. McGraw-Hill Retrieved from http://teachingtoday.glencoe.com/howtoarticles/alternative-assessment-primer.)

*Portfolio Assessment*: A portfolio is a collection of student work that can exhibit a student's efforts, progress, and achievements in various areas of the curriculum. A portfolio assessment can be an examination of student-selected samples of work experiences and documents related to outcomes being assessed, and it can address and support progress toward achieving academic goals, including student efficacy. Portfolio assessments have been used for large-scale assessment and accountability purposes (e.g., the Vermont and Kentucky statewide assessment systems), for purposes of school-to-work transitions, and for purposes of certification. For example, portfolio assessments are used as part of the National Board for Professional Teaching Standards assessment of expert teachers. (Retrieved from answers.com http://www.answers.com/topic/portfolio-assessment.)

*Performance Appraisal*: Performance appraisal is the procuring, analyzing and documenting of facts and information about an employee's net worth to the organization. It aims at measuring and constantly improving the employee's present performance and tapping the future potential.

*Performance Evaluation*: Performance evaluations are prepared by company management on a periodic basis to determine if employees are working up to, or beyond, the minimum standards of their job description. Critical areas are graded by supervisors or department managers in either a written or checklist format, or a combination of both. Decisions ranging from salary increases to possible termination can result from performance evaluations.

*Performance Task*: A performance task is a goal-directed assessment exercise. It consists of an activity or assignment that is completed by the student and then judged by the teacher or other evaluator on the basis of specific performance criteria.

## Current Status

Extensive work on performance testing has been going on since at least the 1960s (e.g., Glaser, 1963; Glaser & Klaus, 1962). Perhaps the best overview of the state of research and practice in educational assessment was given in a recent report on educational assessment by the National Research Council (Pellegrino, Chudowsky, & Glaser, 2001).

In addition there are modern standards for the practice of assessment. The Joint Committee on Standards for Educational Evaluation was formed in 1975 by major

professional associations in social, psychological, and education science and practice. Three sets of standards have been published: Personnel Evaluation (Joint Committee, 1988, revised 2009), Program Evaluation (Yarbrough, Shulha, Hopson, & Caruthers, 2011), and Student Evaluation (Joint Committee on Standards for Educational Evaluation, 2003). In addition, the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1999) have produced standards for educational and psychological testing that cover validity, reliability and error, test development, scoring, score comparability, fairness, and testing applications.

Today there is continuing interest in the role and purpose of testing, and performance-based assessments are in vogue. The US Department of Education's *Race to the Top* assessment program provides "funding for the development of new assessment systems that measure student knowledge and skills against a common set of college and career-ready standards … in mathematics and English language arts in a way that covers the full range of those standards, elicits complex student demonstrations or applications of knowledge and skills as appropriate, and provides an accurate measure of student achievement across the full performance continuum and an accurate measure of student growth over a full academic year or course." (Department of Education, 2010).

Performance assessments often appeal to those who are uncomfortable with large-scale, high-stakes, standardized, normative testing. However, performance assessments have obvious shortcomings. These include: difficulty and expense of administration, unreliability of scoring because of human rater error or bias, and poorer validity or generalizability due to limited time and opportunity to sample extensively from a broad universe of knowledge and skill.

Many performance assessments used in classrooms are teacher designed. This has both advantages and disadvantages. Teachers can adapt their assessments more rapidly to individuals, and can therefore diagnose and remediate more effectively. But this flexibility comes at the price of standardization and reliability, because teachers are often not trained or expert in the design of reliable and valid assessments, and because adaptation necessarily involves alteration of the assessment situation from individual to individual. In addition, teachers may have more difficulty interpreting results that are not expected.

There are more subtle problems in the design of reliable and valid performance assessments. Most performance assessments have associated "rubrics" or scoring keys, which are necessary to provide some standardization and reliability in scoring. Many examples are online at, for example http://www.rcampus.com/indexrubric.cfm or http://www.rubrics4teachers.com/. But in reality, most of these are nothing more than checklists for critical events, and "behaviorally anchored rating scales" from 50 years ago (Smith & Kendall, 1963). At the time these rating techniques were originally developed, such checklists and behavioral anchors depended on reasonably extensive critical-incident and behavioral task analyses. Such analyses have their own difficulties: they are expensive and time consuming to conduct; they depend on deep subject-matter or content knowledge on the part of the analyst; they may be incomplete or inaccurate; they often lead to oversimplification of content (and therefore, for performance measurement, to assessment at too rudimentary a level); and to an overemphasis on procedural skill (rather than underlying cognition) (Bell, Andrews, & Wulfeck, 2010). Ironically, without deep task analysis, performance assessment may be just as behaviorally trivial as poorly designed machine-scored objective tests.

In the end, performance assessments, by definition, rely on some observation of behavior, with inferences about underlying knowledge and cognition. The same is true of any objective assessment which relies on the same sort of behavioral and cognitive task analyses. Therefore, claims concerning the supposed superiority of performance assessment (compared to objective tests) in contemporary education advocacy must be taken with a large dose of skepticism. Indeed, we often find recommendations and practices that, at best, show little awareness of decades of development, and at worst, are simply wrong. This is true not only in the popular press, where policies and practices of education and assessment are often the subject of inflammatory but ill-informed political debate, but also in professional guidance for educators. For example, a publication from a state office of education giving recommendations for science teachers about development of performance assessments says: "…objectively scored tests are not valid measures of what is important to learn in school. Objectively scored tests—multiple choice, completion, short answer—emphasize the acquisition of and the memorization of information. They cannot be appropriately used to measure many higher level thinking abilities nor can they be used to measure some other important goals of schooling." (Baird, 1997).

Leaving aside the bizarre implication that acquisition and memorization of information are not important outcomes of schooling, we concentrate on the claim that objectively scored tests cannot appropriately measure higher-level thinking: While it is true that poorly designed tests might not assess very well, it is certainly not true that assessments cannot be designed to provide objective scoring of many kinds of complex performances, including those that depend on reasoning and problem solving. We have known for decades how to do it, when the tasks for which performance is to be assessed are deeply and explicitly analyzed (e.g., Ellis & Wulfeck, 1982; Glaser & Klaus, 1962; Merrill, 1994; Stevens & Collins, 1977).

We are indebted to a reviewer of this chapter for noting that various forms of assessment cover a broad range of

alternatives, such as the nature of the contextual setting for response or performance, process vs. product measurement, response modes, types of "items" (or instances of response observation), and methods of scoring. Different combinations of alternatives on these dimensions may be more or less appropriate depending on the specific task(s) to be assessed but there should be no *a priori* claim of superior reliability or validity for any particular combination, assuming, of course, that the assessment is competently designed. Further, however, the choice of different combinations may have substantial implications for the cost and practicality of development, administration, scoring, reporting, and utilization of assessment results. For example, "paper" simulations with "multiple-choice" responses can be designed for many types of mechanical or electronic troubleshooting tasks: they can test logical reasoning and problem solving and can provide diagnostic information concerning misconceptions (cf. Ellis & Wulfeck, 1982). They are much cheaper to administer and score than an actual hands-on troubleshooting event that requires live test equipment and individual administration and scoring by a human observer.

## Simulation-Based Performance Assessment

Even before the large-scale development of computer-based simulations, case-based and role-based simulations were used in education, such as "moot courts" in legal training, or wargaming in the military. Non-computer-based simulators are heavily used in many fields, for example, in medical education (cf. www.simulation.com for examples of commercially available human patient simulators), or in firefighting training (cf. www.mobilefireunits.com for examples of commercially available simulators).

In simulation-based training, the simulation is used to provide the context in which human performance may occur and be observed, recorded and measured. The most important ingredients for successful simulation-based training and assessment are the design of the scenarios, since these prescribe the conditions under which performance will be elicited, the physical fidelity of the simulation (for example in medical devices), and (as in any performance assessment) the scoring criteria against which performance will be evaluated. In simulations which are not computer based, observation and measurement generally use the same techniques as in other performance assessments, namely checklists, rating scales, and occasionally time-to-solution measures. These, of course, suffer from the same limitations as most other performance assessments. For example, a review of assessments used in anesthesia simulation in medical training found few which addressed questions of validity or reliability (Byrne & Greaves, 2001).

## So, What's New in Assessment?

In recent years, with the development of computing technologies, it has become possible to build detailed, highly veridical simulations of complex phenomena (cf. Baker, Dickieson, Wulfeck, & O'Neil, 2008). Computer-based simulations are now used in many fields of endeavor, such as medicine and surgery, engineering, economics, geology, vehicle piloting, and many others. Simulations are used for system design, system performance analysis, prediction of outcomes, analysis of alternative courses of action, and of course for training and performance assessment.

Computer-based simulations have some properties which may contribute to effective performance measurement. First, the process of constructing the simulation essentially involves a very fine-grained task analysis, since almost every aspect of the user's interaction with the simulation must be taken into account in its design (Wulfeck, Wetzel-Smith, & Dickieson, 2004). Thus, unlike performance assessments where variations in performance can be handled (or missed) by a human rater's observational skill, such variation must be explicitly accounted for in the simulation. Second, the simulation can be designed to collect performance data automatically, and use it for scoring and evaluation, as well as for control and adaptation in the simulation itself. Third, the simulation is, by definition, situated in a task environment, so criticisms concerning the unreality of common standardized multiple-choice testing are avoided.

## Simulation-Based Performance Assessment for Aviation

Before we discuss simulation-based performance assessment for aviation it is important to discuss simulation fidelity in its various forms. There are a number of different types of simulation fidelity (Hays & Singer, 1989; Swezey & Andrews, 2001). Physical fidelity refers to the physical characteristics of the simulation or simulator ("does it look right?"). Functional fidelity refers to the way the simulation or simulator behaves ("does it act right?"). Psychological fidelity, which is much more difficult to assess, refers to how an experienced real-world operator of a system believes the simulation or simulator subjectively meets their expectations in terms of its "feel" to them ("does it feel right?"). For example, an aircraft simulator may look and act like the real thing, but an operator may still not get an authentic feeling when they fly the simulator. "Cognitive fidelity" is a construct similar to psychological fidelity.

A simulation or simulator might have different levels (low to high) of each type of fidelity. The training developer and/or assessment developer must decide about the optimal mix of

fidelity levels for the simulation's intended purpose. The cost to achieve higher levels of fidelity and the technical challenges in reaching higher levels both enter into the decision. The authors have seen many instances where large amounts of money have been invested to achieve high levels of physical and functional fidelity and yet the operational experts still did not feel that the simulation felt like the real thing. It did not reach a high level of cognitive fidelity. Yet, we have also seen example where physical/cognitive fidelity was achieved with relatively modest investments in physical/functional fidelity. The key is determining as precisely as possible what salient characteristics of the real-world system must be represented in the simulation in order for learning to occur, or in order for the trainee to be able demonstrate competence to an assessor.

A key example of the use of simulation to enable assessment comes to us from years of research and practice in the pilot assessment arena. From the beginning of pilot training, concern has been given to how best to measure pilot performance (Meister, 1999). Assessment of this type is used both in training and in assessing readiness for job performance. Pilots must make rapid decisions in highly dynamic and complex environments. Since the time of the Wright Brothers, the main method for assessing pilot trainee and pilot performance has been via rating sheets, typically with five or seven point rating scales. The instructor pilot, flying with the trainee, determines the pilot's performance for each of the scale dimensions (e.g., mission planning, preflight check, taxi, takeoff, aerial maneuvers, stall recovery, instrument flight, situational awareness, etc.). Once a military pilot progresses beyond their initial undergraduate training and is ready for combat training, the instructor pilot assesses their performance on various phases of tactical flying. This happens not only in their initial aircraft specific training but also periodically throughout their flying career.

This approach for assessing trainee and pilot performance is a tried and true method. Although ultimately subjective in nature, it has proven to be generally valid and reliable. Instructor pilots have come up through this type of assessment system as they gained expertise and are quite used to making judgments on the rating scales after observing trainee/pilot performance.

Andrews, Nullmeyer, Good, and Fitzgerald (2008) provide an overview of the progress made in performance measurement in the aviation field. They note two advancements that deserve highlighting. The first is the advent and use of digital performance recording as described above. This has led to new approaches to automated performance assessment, as well as giving instructors new tools to make subjective assessment decisions. The second advancement is the use of behaviorally anchored rating scales coupled with automated performance assessment tools to develop a more robust total measurement system.

Aircraft simulators have opened up a broad new horizon for performance assessment in pilot training. Digital simulators can record every aspect of the simulated aircraft on a micro-second basis. Every button push, toggle switch movement, image display, radar mode, etc. can be recorded for eventual analysis by instructors and raters. This data can be aggregated both for individual trainee evaluation and across trainees to spot gross trends that can lead to training program improvements. In addition, all audio communication can be recorded and replayed.

In many cases, measurement done in a simulator can be more valid and reliable than measurement done by an instructor pilot. That is especially true in the early phases of training. If the simulator possesses both high physical and functional fidelity it is a relatively straight forward task to measure each button push and control movement. The validity and reliability become more difficult to maintain as the pilot moves into phases of flight that demand more cognitive skill (e.g., understanding how to synthesize the information from various displays, decision making, etc.). In those cases the automated performance measurement system can aid the instructor, but normally can't provide all of the measurement capability required.

One potential advantage of simulator-based assessment systems is that they may give the trainee the feeling that they are not constantly being watched by their human instructor. A number of researchers (Diaper, 1990; Shivers, 1998; Staal, 2004) have shown that performance is altered when it is done in the presence of others. Performers who were unaware that they were under observation typically performed better than a control group on complex tasks. One might then conclude that having an automated performance system do the assessing improves performance if the instructor is not directly involved with observing the trainee's performance.

Lane (1986), in a report examining the performance measurement issue in aviation, cites "seven" criteria that can be used in evaluating aviation performance measurement. They are as follows:

*Reliability*—"Reliability … is in the metric sense the most basic issue. If the measures are not dependably replicable over the required time period, other criteria are of little or no importance." p. 36

*Validity*—Lane provides a cogent quotation from Wallace (1965) that explains the importance of true validity, "… it is possible to develop extremely plausible measure sets, with high apparent relevance, which are in reality mostly irrelevant and provide no evidence of any sort germane to the purpose of the evaluation."

*Sensitivity*—"The sensitivity of a measure reflects the extent to which the measure behaves 'appropriately' in response to changes in conditions under which the task is performed or to differences in individual capability to do the task." p. 80

*Completeness* (*Dimensionality*, *Comprehensiveness*)—"… basic, advanced and operational flying, evaluators made consistent and reliable distinctions between such aspects of proficiency as basic airwork, instrument flying and ability to use weapons." p. 81

*Seperability of Operator from Measurement Context Contributions*—"If comprehensiveness is the inclusion of all the relevant components of performance, then the concern for seperability is for the omission or exclusion of irrelevant components." p. 91

*Diagnosticity* (*Specificity*)—"To be effective in … diagnostic use, variables must satisfy three general requirements: a) they must provide a level of detail which allows differentiation among skill and knowledge components, b) they must be sufficiently distinct in the content they measure, and c) the measures must be capable of being mapped with a reasonable degree of correspondence into those specific components." p. 93

*Utility and Cost Benefit* (*Value against Alternatives*)—"A measurement system may be reliable and valid and possess all the other properties required of performance measures and still be of limited utility … To be 'useful', a method must produce results that represent 'true' performance more closely than any other available and affordable way of achieving that objective." p. 94.

Lane's (1986) comments concerning performance assessment in aviation apply well to most new applications of performance assessment technology. While some progress has been made in recent years in applying more modern statistical approaches to questions concerning reliability, validity, and generalizability of performance measurements (cf. Webb, Shavelson, & Haertel, 2006), much remains to be done to build an engineering science using modern computer-based simulations for performance assessment.

## Non-aviation Example of Simulation-Based Performance Assessment

An example might help the reader to understand how a simulation-based approach might be of use in K-16 schooling and in higher education. In a history course a teacher wants students to develop more complex decision making capacity. They wish to have the students exercise their abilities to define a problem, describe the key elements of a decision and solution requirements in order to frame their solution alternatives, pick a solution strategy and apply it, evaluate the results of the decision and make revisions where necessary or reject that alternative and choose a different solution strategy. They could use an off-the-shelf game that re-creates a financial emergency such as the great depression. It is not likely that an off-the-shelf game would have the kinds of assessment characteristics that a teacher would likely desire, so they would need to conduct the assessment of student skills manually, but the main point is that the students would be developing decision making skills in a real-world simulation with the types of variability that would expect to be found in the actual historical settings. The huge success enjoyed by decision making games using real-world simulation, (e.g., SimCity and Age of Empires) shows that schoolchildren enjoy such games. We highly encourage educational game makers (i.e., serious games) to build in the types of assessment characteristics discussed in this chapter.

Assessment of medical skills is another area that is benefitting from simulation-based education and training. Past medical specialty certifications by national boards for that discipline consisted primarily of written exams and interviews between the applicant and experts in that field. The expert would ask various questions about the specialty and often pose relevant questions about best treatment paths using verbal scenarios. While that approach is still in use in some cases, the certification boards are turning more and more to simulation-based assessment approaches that can supplement the other elements of the assessment process. These computer based assessment systems allow the certifying assessors to test the medical specialist applicants' hands-on skills. This trend will continue to gain momentum.

## Future Research Needs

A fruitful area of research involves extending simulation design to the measurement and assessment of high-level cognition. An entire volume on assessing problem solving using simulations explicitly treated issues and examples in the design of simulations for assessment (Baker et al., 2008).

A second area of research and development would extend the progress made in the use of simulation for assessment for broader application, particularly in education at the preschool through college levels. Here the need is primarily for diagnosis of progress and prescription of curricular and instructional alternatives that maximize the progress of individual students (cf. Lesgold, 2008). These areas of research have been seriously neglected in education and psychology in recent years, due to intense pressure from politicians and others who apparently think that a focus on scores on standardized tests will somehow improve competitiveness in a world economy. While a quick search of the Internet will locate hundreds of computer-based simulations or environments for K-12 tasks, few have mechanisms for recording student interaction to permit significant assessment or diagnosis. This is due in part to the relative infancy of computer applications in education, and in part to the difficult problems of practical management of such systems in today's classrooms.

A third related area of research involves performance assessment for high-level tasks which are complex and ill-structured. These are often the highest-value tasks/skills at

any level of work, in any kind of organization. Wulfeck and Wetzel-Smith (2008) described these "Incredibly Complex Tasks" as those that are almost unbelievably complicated, in that they required deep expertise obtained through years of highly contextualized study, practice, and experience for successful performance. In later work, Wulfeck and Wetzel-Smith (2010) described training strategies for incredibly complex tasks. These involve using computer modeling, visualization, and careful design of instruction to deal with task characteristics that contribute to difficulty and complexity, such as multiple sources of variation, interaction, dynamism, continuity, nonlinearity, simultaneity, uncertainty, and ambiguity. These design strategies can also inform the design of performance assessments. Unfortunately, however, developing training and performance assessments for incredibly complex tasks requires at least as high a level of expertise as performance of the tasks themselves, and so such tasks are often considered by non-experts (such as policy-makers) to be impossible to teach or to assess. On the contrary, it is entirely possible to design good assessments for incredibly complex tasks: it merely requires decades of work by task experts and millions of dollars. However, efforts on such a scale, even in education and training, are common. Standardized testing programs for *No Child Left Behind* or college entrance examinations have already consumed much larger amounts of resources.

## Conclusions

After a combined 80 years as practicing instructional psychologists, we welcomed the opportunity to examine current developments in the area of performance assessment. In the past 20 years there has been tremendous progress in the application of technology to performance assessment, particularly in developing and applying simulation systems for both individual and team training. Much of this progress has come in the military in the development of simulation-based training systems, and in some professions for certification.

Performance assessment has become more important over time. In our litigious society, the capability to accurately determine who can do what not only makes good educational and economic sense but it also can protect organizations from being sued for hiring, promotion and termination decisions. As a scientific community and community of practice, our performance-assessment standards and tools continue to evolve. We are considerably more likely to make the right judgments about personnel and student performances now than we were, say, 50 years ago. Not only have the accuracy and reliability of our assessments improved, but also the standards and tools we possess make us better able to anticipate and react to, and in some cases mitigate, both individual and societal differences. In addition, our assessments better help our community to make sound prescriptions to improve performance for students and employees.

Despite the progress in assessment over the last few decades considerable work remains to be completed. Research into performance assessment techniques that translate from tests to real-world performance is still inadequately funded. At least some of the considerable public debate about the place of testing in schools stems from a general misunderstanding, and in some cases mistrust, of validity of the assessment process at measuring true educational progress. While there is a general sense that the standardized tests used to measure that progress are not complete, in some cases there is the feeling that the tests are biased in one way or another. Researchers in the educational measurement community have work to do in establishing validated methods for identifying key concepts and principles required in the various academic subjects that can be translated into effective tests.

We generally believe it is a more straightforward, although by no means simple, task for measurement specialists in business, industrial and military settings to develop quality measurement approaches than it might be in formal education. Proximity of their assessment development activities to real-world work places gives them an advantage over their formal education counterparts. Having completed a number of task analyses, from which flow metrics for assessment, your authors know the advantage that comes from being able to observe and interview incumbent workers as they perform their jobs. Formal education assessment developers have access to subject matter experts, but seldom get to watch them use their expertise to accomplish real-world tasks, whether that is performing cognitive or manual work tasks.

However, the task of workplace performance assessment is made difficult because of the increasing complexity of many jobs. While automation has simplified many workplace tasks it has also forced decisions about the allocation of job functions that have resulted in humans taking on more executive control functions. These are almost always more difficult to assess than jobs where the human is doing more procedural work. While the mundane parts of jobs are performed by sophisticated software, the human is left to monitor the job activity and decide when to intervene in the process. These control decision tasks were usually left to senior, more experienced, and better trained employees, but they have been often pushed down to newer and less trained employees. Designing performance measurement systems can help in assessing the preparation of employees who will make these decisions. Since in some cases there may not be only one right answer in a workplace setting, performance assessment systems will have to increase in flexibility.

**Acknowledgments** Portions of this work were supported by the US Navy. The views and opinions expressed herein are those of the authors, and are not to be construed as official or as representing the Department of Defense, or the Department of the Navy. In addition, we thank two anonymous reviewers for their helpful comments that improved this manuscript.

## References


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

*Andrews, D., Nullmeyer, R., Good, J., & Fitzgerald, P. (2008). Measurement of learning processes in pilot simulation. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil. *Assessment of problem solving using simulations* (pp. 273–288). New York: Lawrence Erlbaum Associates.

Anderson, J. R., Reder, L. M., & Simon, H. A. (2000). Applications and misapplications of cognitive psychology to mathematics education. *Texas Educational Review, 6*.

*Baker, E. L., Dickieson, J. L., Wulfeck, W. H., & O'Neil, H. F. (Eds.). (2008). *Assessment of problem solving using simulations.* New York: Lawrence Erlbaum Associates.

*Bell, H. H., Andrews, D. H., & Wulfeck, W. H., II. (2010). Behavioral task analysis. In K. H. Silber & W. R. Foshay (Eds.), *Handbook of improving performance in the workplace, vol. 1: Instructional design and training delivery* (pp. 184–226). San Francisco, CA: Pfeiffer/International Society for Performance Improvement.

Baird, H. (1997). *Performance assessment for science teachers*. Salt Lake City, UT: Utah State Office of Education.

Bloom, S. B. (1956). *Taxonomy of educational objectives*. Boston: Allyn and Bacon.

Byrne, A. J., & Greaves, J. D. (2001). Assessment instruments used during anesthetic simulation: Review of published studies. *British Journal of Anaesthesia, 86*(3), 445–450.

Department of Education. (2010). *Overview information*; *Race to the top fund assessment program*; *Notice inviting applications for new awards for fiscal year* (*FY*) *2010*. Federal Register/Vol. 75, No. 68/ Friday, April 9, 2010/Notices. p. 18171.

Diaper, G. (1990). The Hawthorne effect: A fresh examination. *Educational Studies, 16*, 261–267.

Ellis, J. A., & Wulfeck, W. H. (1982). *Handbook for testing in Navy schools*. Special Report 83-2, San Diego, CA: Navy Personnel Research and Development Center. DTIC Accession Number: ADA122479.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performances. In R. M. Gagné (Ed.), *Psychological principles in systems development* (pp. 419–474). New York: Holt, Rinehart, & Winston.

Hays, R. T., & Singer, M. J. (1989). *Simulation fidelity in training system design: Bridging the gap between reality and training*. New York: Springer-Verlag.

Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards: How to assess systems for evaluating educators*. Newbury Park, CA: Sage Publications.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards: How to improve evaluations of students*. Newbury Park, CA: Corwin Press.

Lane, N. E. (1986) *Issues in performance measurement for military aviation with applications to air combat maneuvering*. Technical Report: NTSC TR-86-008. Naval Training Systems Center. DTIC Accession Number: ADA172986.

Lesgold, A. (2008). Assessment to steer the course of learning. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 19–36). New York: Lawrence Erlbaum Associates.

Madaus, G. F., & O'Dwyer, L. M. (1999). Short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688–695.

Meister, D. (1999). Measurement in aviation systems. In D. J. Garland, J. A. Wise, & V. D. Hopkins (Eds.), *Handbook of aviation human factors* (pp. 34–49). Mahwah, NJ: Lawerence Erlbaum Associates.

Merrill, M. D. (1994). *Instructional design theory*. Englewood Cliffs, NJ: Educational Technology Publications.

*Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Shivers, C. H. (1998). Halos, horns and Hawthorne: Potential flaws in the evaluation process. *Professional Safety, 43*(3), 38–41.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors to rating scales. *Journal of Applied Psychology, 47*, 149–155.

Staal, M. A. (2004). *Stress, cognition, and human performance: A literature review and conceptual framework.* NASA Tech. Rep. NASA/TM-2004-212824. Moffat Field, CA: Ames Research Center, Retrieved from http://human-factors.arc.nasa.gov/flightcognition/Publications/IH_054_Staal.pdf

Stevens, A., & Collins, A. (1977). The goal structure of a Socratic tutor. *Proceedings of Association for Computing Machinery National Conference.* Seattle, Washington.

Swezey, R. W., & Andrews, D. H. (2001). *Readings in training and simulation: A 30-year perspective*. Santa Monica, CA: Human Factors and Ergonomics Society.

U.S. Congress, Office of Technology Assessment. (1992). *Testing in american schools: Asking the right questions, OTA-SET-519*. Washington, DC: U.S. Government Printing Office.

Wallace, S. R. (1965). Criteria for what? *American Psychologist, 20*, 411–417.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao (Ed.), *Handbook of statistics* (Volume on Psychometrics, Vol. 26, pp. 81–124). Amsterdam, The Netherlands: Elsevier.

Wiggins, G. P. (1993). *Assessing student performance*. San Francisco, CA: Jossey-Bass Publishers.

Wulfeck, W. H., & Wetzel-Smith, S. K. (2008). Use of visualization to improve high-stakes problem solving. In E. L. Baker, J. Dickieson, W. H. Wulfeck, & H. F. O'Neal (Eds.), *Assessment of problem solving using simulations* (pp. 223–238). New York: Lawrence Erlbaum Associates.

Wulfeck, W. H., & Wetzel-Smith, S. K. (2010). Training incredibly complex tasks. In P. E. O'Connor & J. V. Cohn (Eds.), *Human performance enhancement in high risk environments*. Santa Barbara, CA: ABC-CLIO.

Wulfeck, W. H., Wetzel-Smith, S. K., & Dickieson, J. L. (2004). Interactive multisensory analysis training. In NATO RTO Human Factors and Medicine Panel (Eds.), *Symposium on advanced technologies for military training*. Neuilly-sur-Sein Cedex, France: North Atlantic Treaty Organization Research and Technology Agency.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards* (3rd ed.). Thousand Oaks, CA: Sage.