RESEARCH BRIEF

Institute for Homeland Security Solutions
Applied research • Focused results

April 2009

Job Proficiency Work Sample Testing for Critical DHS Jobs and Job Tasks

Project Leads

Jerry Hedge, PhD, RTI International

Robert Hubal, PhD, RTI International

Statement of Problem

The Department of Homeland Security works to anticipate, preempt, detect, and deter threats to the homeland. Consequently, vigilance is a primary aspect of many "front-line" jobs oriented toward detection and prevention, where employees must pay close and sustained attention and maintain that attentiveness over time. Often this is to be found in some form of "watchkeeping" activity when an observer, or listener, must continuously monitor a situation in which significant, but usually infrequent and unpredictable, events may occur. A great deal of emphasis is necessarily placed on correctly detecting the presence of inappropriate events and dismissing the presence of appropriate events. Viewed another way, considerable time is spent not seeing anything and not finding anything.

From a measurement perspective, gauging the performance levels of these employees may be particularly problematic. While traditional performance management and performance appraisal systems can be effective at capturing individual performance levels across jobs and across work units (see, for example, Hedge, Borman, Bruskiewicz, & Bourne, 2004), such systems rely on overall impressions, or the observation that an individual is responding

properly and at required performance levels on *routinely observable* aspects of the job. These approaches would tend to miss or underrate activities that are important but less frequently performed. As a result, critical job components are overlooked—in terms of evaluating and maintaining job proficiency levels. In spite of these problems, such performance management systems are often utilized by organizations (particularly large and dispersed organizations) because they offer an expedient means of gathering performance information across an entire organization.

Other organizations, in part concerned about the subjectivity of ratings, prefer to rely instead on so called "hard" criteria (number of apprehensions, time required to respond to an emergency situation, etc.) because they are believed to represent more objective performance data. Unfortunately, these approaches also have their weaknesses. For example, one measurement metric used by the U.S. Department of Homeland Security's (DHS) Bureau of Customs and Border Protection (CBP) to gauge the performance effectiveness of their Air and Marine, Border Patrol, and Field Operations units is an *absolute count of the number of drug seizures per evaluation period*. DHS's Office of Inspector General (2009) recently released a Drug Control Performance Summary Report that noted CBP's opposition to setting drug seizure performance targets and emphasized that such targets can be misleading or counterproductive. The report concluded that such performance targets alone may offer an incorrect assessment of agency success. Similarly, a recent Congressional Research Service (CRS) Report for Congress on the role of the U.S. Border Patrol (Nunez-Neto, 2008) suggested that an overemphasis on use of apprehension statistics as a performance measure may be misleading because, for example, successful illegal entries do not get tabulated, and therefore, such data may represent a fairly unreliable gauge of attempted entries.

In other words, a major weakness of relying exclusively on such objective measures is that they tend to be *deficient* in terms of adequate coverage of the performance domain. When organizations want or need to examine individual competencies at a more detailed level, gauge performance on critical but infrequently performed tasks, or establish and maintain minimum competency performance levels for particularly crucial tasks or activities, traditional approaches prove much less useful, and other techniques must be adopted. This research brief examines one such approach (simulation environments for work sample testing) that DHS can employ to identify task-specific competency levels within key jobs and ensure that proficiency is maintained at high levels for these critical job components.

Background

Organizations routinely strive to gauge the performance levels of their employees and seek to gain insights into overall organizational productivity and knowledge and skill levels of individual workers. Job performance is a complex concept that can be measured with a variety of techniques, including rating scales, job knowledge tests, and work sample tests. Because of

their *relative* ease of development and adaptability to different jobs, rating scales employed in conjunction with observations are used most frequently in organizational settings, but as noted, they have their limitations. Another frequently used approach is job knowledge tests, which are easy to administer but require the individual to demonstrate knowledge *about* the skills under study, as opposed to demonstrating the skills directly.

Historically, work samples or performance tests have been used more commonly as predictors than as criterion measures. However, as Hedge and Borman (1995) noted, in one sense they do represent a compelling criterion measurement method: What could be a fairer measure of job performance than to ask individuals to complete several of the most important tasks that comprise their job, and then evaluate their performance on those tasks? In fact, a number of researchers (e.g., Ghiselli & Brown, 1948; Guion, 1979; Robertson & Kandola, 1982; Siegel, 1986) have advocated the usefulness of the work sample methodology because it affords direct, relevant measurement of job proficiency by extracting samples of behavior under realistic job conditions and asking individuals to demonstrate proficiency on those tasks by performing (in a hands-on manner) the steps required for successful completion.

The situation requires the individual to demonstrate mastery of task performance in a somewhat realistic setting and, by implication, possession of the requisite skill levels required for actual performance on the job. Every individual performs the same task(s) under the same conditions and, as allowed by task conditions, is scored in a standardized way. Past research has shown that work sample tests display high content validity and face validity (e.g., Guion, 1998; Hedge & Teachout, 1992). In addition, the test content's direct relevance to the job may ensure that applicants and assessors view the tests more favorably.

Work sample tests have been used for many years as a method for *selecting* applicants into the workforce. As such they have been designed primarily to assess present skill levels. A related use of work sample testing as a selection tool is exemplified by the research of Robertson and Downs (Downs, 1970; Robertson & Downs, 1979, 1989) and Siegel (Siegel, 1983; Siegel & Bergman, 1975). This approach, referred to as trainability testing, or miniature job training and evaluation, focuses on identifying an individual's potential for training prior to being placed on the job. Job applicants are given short training sessions followed by testing sessions that assess what has been learned. The success of this approach has been reported by Robertson and Kandola (1982) and Robertson and Downs (1989).

Whereas personnel selection has been the primary reason for work sample test development, work sample tests have also been used to (1) evaluate the worth of training programs (i.e., determine that the instructional program was responsible for the changes that occurred; Goldstein, 1974); (2) identify individual strengths and weaknesses to determine the skills and knowledge needed to perform the job successfully (Goldstein, 1980); and (3) certify that an individual meets or exceeds a designated level of competence at performing a job (Guion, 1979).

However, although use of a hands-on work sample approach has proven effective in a variety of situations and for a variety of purposes, such hands-on testing can be problematic because of the complexity and expense involved in performing many tasks. In addition, it may not be feasible to use the hands-on methodology to measure performance on some critical tasks that may take too long to complete, require replacement of expensive parts, or risk possible injury to personnel or damage to equipment. Additionally, more relevant to DHS, these tasks may also involve infrequent but critical events, as is the case with watchkeeping tasks. Consequently, alternative work sample measurement approaches have been examined.

For example, Hedge and Teachout (1992) developed an innovative approach known as Walk-Through Performance Testing (WTPT), which was designed to overcome some of these same types of situations in an Air Force context. The WTPT was designed to integrate a hands-on and interview approach to criterion development within a work sample framework. The WTPT has as its foundation the work sample philosophy, but it expands the measurement of critical tasks through the use of an interview testing component to include tasks not measured by hands-on testing. The administrator assesses an incumbent's proficiency on a task by evaluating how well his or her show-and-tell descriptions reflect proficiency-based strengths and weaknesses related to the performance of that task.

With the advent of more sophisticated technologies for presenting realistic job samples and scenarios and for capturing performance in those situations, use of a work sample approach to measuring job proficiency has become more viable. For example, a complementary line of research aims to place the individual within a *simulated* environment that closely mirrors the real environment and assess the individual's performance. The situation might measure physical, tactile, verbal, interactive, or other behavior. Another potential benefit of a simulation approach to performance measurement might include its application to a much broader population. Because of the prevalence of computer technology in most workplaces, simulated performance scenarios can be adapted for use within a large and nationally dispersed workforce such as DHS. For example, measuring the job performance of air traffic controllers, among other occupations, is a situation where reliance on such a work sample methodology may be especially applicable. The likelihood of accurately measuring controller performance increases to the extent that one is able to place controllers in a standardized and realistic environment in which they must control traffic and which affords reliable measurement of their performance.

Hedge and colleagues (1997) used a computer-generated simulation to create such an air traffic control environment. The dynamic simulation allowed the controller to direct the activities of a sample of simulated air traffic, performing characteristic functions such as ordering changes in aircraft speed or flight path, but within a relatively standardized work sample framework. In addition, performance data were captured by means of behavior-based rating scales and checklists, using controllers with considerable air traffic experience as raters. A subset of the controlling actions taken by the air traffic controllers during the scenarios was

also captured via computer. This simulation approach proved to be particularly effective and was also used as a high fidelity criterion measure against which to validate other lower fidelity but more readily transportable performance measures.

Another view of how simulation can influence performance assessment is to consider different types of job-related skills. For instance, maintenance training systems typically employ a combination of desktop training devices and hands-on trainers and as such address hands-on or procedural skills. The assessment or validation of learned procedural skills within simulation systems has been shown to be a cost-effective solution (Hubal, 2005; McMaster et al., 2002). Similarly, operation and setup of complex systems, such as signal and communication systems (Evens, Whiteford, Frank, & Hubal, 2006), require that individuals understand not just the procedures but strategies that inform procedures (such as aspects of good site selection) and the need for adaptive responses (Krizowsky, Waters, Wright, Hubal, & Frank, 2008). In all of these systems, the individuals must perform all of the steps that they would perform in the real world, within faulted and unfaulted modes, while instructors can observe the outcomes of their actions and assess their demonstrated strategic proficiency.

These procedural and strategic skills can be assessed in other work domains that more closely resemble watchkeeping activities. For instance, “first responders” are called upon to deliver pre-hospital, primary, and emergency medical care. Simulations that are developed to assess medical first response behaviors (Kizakevich et al., 2001, 2006) offer realistic practice by presenting a scenario comprising an interactive scene, an incident that produces trauma or medical conditions, and one or more simulated patients. The caregiver needs to navigate and survey the scene, interact and converse with the patients, use medical devices, administer medications, monitor diagnostic data, and perform interventions. This work relates directly to DHS occupations involving dispatch and first response.

In contrast to procedural skills, a line of research has investigated the use of responsive virtual characters for critical “soft” skills that are required during an interview, a negotiation, an interrogation, and a de-escalation, such as typically occur during passenger screening and apprehension of suspicious persons. Example applications using this technology include those for training police officers handling encounters with mentally disturbed individuals (Frank et al., 2002), clinicians interviewing either pediatric patients (Hubal, Deterding, Frank, Schwetzke, & Kizakevich, 2003) or patients exposed to bioterrorist agents (Kizakevich et al., 2004), and emergency response personnel encountering trauma patients (Kizakevich et al., 2001) and for assessing the social competency of at-risk students and prisoners (Hubal et al., 2008). These soft skills, then, may play an important role in the successful performance of many DHS-related watchkeeping tasks as well.

Synthesis

For organizations to maintain a highly proficient work force, they must be able to accurately measure individual performance levels and identify skill deficiencies at the task level. Unfortunately, traditional performance management approaches that emphasize the rating of broad competencies prove much less useful. When organizations want or need to examine individual competencies at this more detailed level, or when situations demand establishing and maintaining minimum competency performance levels for particularly crucial tasks or activities, an alternative approach is necessary. For DHS, this is particularly true for front-line occupations and critical tasks within those occupations. Emergency responders, jobs within the Bureau of Customs and Border Protection, and aviation security positions all fulfill particularly crucial roles to anticipate, preempt, detect, and deter threats to homeland security. In addition, the number of employees in these occupations has grown substantially since 2001, increasing both the importance and difficulty in maintaining high performance levels and recognizing problems in sufficient time to successfully intervene.

Past work sample research has demonstrated the usefulness of this measurement approach for establishing and maintaining task-specific performance. In addition, it allows the measurement of proficiencies on tasks (or groups of tasks) that cannot feasibly be measured in a hands-on mode. It might also provide an opportunity to discover whether tasks associated with particular skill sets are more or less sensitive to skill decay. One expected advantage of the simulation technology is that it combines the rigors of work sample testing with measurement efficiency and transportability. In addition, a simulation environment is safe, consistent, and replicable. The development of a reliable and valid simulation measurement approach could prove particularly valuable for DHS.

Future Directions

To capitalize on the application of the work sample job proficiency methodology within a simulated work environment, and to ensure optimal levels of job performance associated with crucial DHS skill sets, we recommend that future research proceed in the following manner. A first critical step would be the design and implementation of a process for identifying jobs that deal with the greatest vulnerabilities to homeland security (and therefore benefit most from a detailed proficiency analysis) and would be most amenable to the simulated work sample methodology.

This requires the use of currently available job analysis information or the collection of these types of data (using standard job analysis methods or cognitive task analysis methodologies) if no such data exist. Such techniques define the job in terms of time spent on a number of discrete and interrelated tasks and their importance to job success. Of particular

interest to DHS will be selection of those tasks and their associated skill sets deemed most critical and especially sensitive to proficiency decay. One such occupation may be border patrol, where sustained vigilance is an important job component. The type of work activities representative of this job would also be well-suited to a simulation environment. A second type of job family that presents a good candidacy would be first responders, who are sometimes called on to perform a set of job activities that may differ considerably from their routine daily job activities. Because they may perform such emergency job activities only rarely, their proficiency levels are subject to performance decay.

Throughout this research brief we have used examples of potentially relevant DHS jobs, such as first responders, aviation security, and border patrol, but many other jobs would benefit greatly from this application, and a methodology that allows a closer look at, and prioritization of, DHS jobs should be quite useful; for example, intelligence analyst positions require extensive “monitoring” and analysis activities.

A second and related research direction that could pay immediate dividends to DHS would be an analysis of products, technologies, and processes that currently exist within DHS and could be utilized to support the simulation effort. For example, FEMA currently oversees some first-responder training protocols that might be amenable to adaptation. In particular, their Homeland Security Exercise and Evaluation Program (HSEEP), designed for training purposes, might be usefully applied to performance measures. A thorough examination of available technologies would identify where simulations would have to be developed and where simulation work could be adapted from exiting technologies.

Together, these components could serve as first steps to establish the foundation for future job proficiency measurement programs and, in a sense, help establish a future roadmap for this research approach. Inherent in such an approach would be a clearer delineation of skill set perishability, and a potential performance-skill decay taxonomy. The longer term potential is application of a technology to ensure high levels of job proficiency for all jobs within the DHS wherein even a momentary lapse in skill levels can lead to dire consequences. Similarly, other training tools should be examined to determine whether they offer cost-effective alternatives or applications for building a work sample methodology. Using this methodology, it would be possible to establish performance protocols and accompanying performance standards for different jobs and skill levels (e.g., apprentice, journeyman, master) to help strengthen the performance capabilities of DHS personnel.

Contact Information

Jerry Hedge, PhD
RTI International
3040 Cornwallis Rd.
Research Triangle Park, NC 27709
(919) 541-6496
jhedge@rti.org

Robert Hubal, PhD
RTI International
3040 Cornwallis Rd.
Research Triangle Park, NC 27709
(919) 541-6045
rhubal@rti.org

Jerry Hedge, PhD, is an industrial/organizational psychologist at RTI, with extensive technical experience in job analysis and performance measurement. He has conducted numerous projects for public and private-sector organizations in these areas, and he has contributed frequently to the research literature on these topics.

Robert Hubal, PhD, is a cognitive scientist at RTI long interested in the intelligent application of novel technologies to pressing training and assessment needs. He studies effectiveness, efficiency, acceptance, and usability of these applications in both everyday and specialized domains.

References

Downs, S. (1970). Predicting training potential. *Personnel Management*, *2,* 26-28.

Evens, N., Whiteford, B., Frank, G., & Hubal, R. (2006). User interface lessons learned from distributed simulations. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference* (pp. 1276-1285). Arlington, VA: National Defense Industrial Association.

Frank, G., Guinn, C., Hubal, R., Pope, P., Stanford, M., & Lamm-Weisel, D. (2002). JUST-TALK: An application of responsive virtual human technology. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference* (pp. 773-779). Arlington, VA: National Defense Industrial Association.

Ghiselli, E., & Brown, C. (1948). *Personnel and industrial psychology.* New York: McGraw-Hill.

Goldstein, I. L. (1974). *Training: Program development and evaluation.* Monterey, CA: Brooks/Cole.

Goldstein, I. L. (1980). Training in work organizations. *Annual Review of Psychology, 31,* 229-272.

Guion, R. M. (1979). *Principles of work sample testing: I. A non-empirical taxonomy of test uses* (ARI-TR-79-A8). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Erlbaum.

Hedge, J. W., & Borman, W. C. (1995). Changing conceptions and practices in performance appraisal. In A. Howard (Ed.), *The changing nature of work* (pp. 451-481). San Francisco: Jossey-Bass.

Hedge, J. W., Borman, W. C., Bruskiewicz, K. T., & Bourne, M. J. (2004). The development of an integrated performance category system for supervisory jobs in the United States Navy. *Military Psychology, 16,* 231-243.

Hedge, J. W., Bruskiewicz, K. T., Manning, C., & Mogilka, H. (1997). *The development of a high fidelity work sample measure of performance for air traffic controllers* (Institute Report # 297). Minneapolis: Personnel Decisions Research Institutes, Inc.

Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion measurement. *Journal of Applied Psychology*, *77*, 453-461.

Hubal, R. (2005). Design and usability of military maintenance skills simulation training systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 2110-2114). Santa Monica, CA: Human Factors and Ergonomics Society.

Hubal, R. C., Deterding, R. R., Frank, G. A., Schwetzke, H. F., & Kizakevich, P. N. (2003). Lessons learned in modeling virtual pediatric patients. In J. D. Westwood, H. M. Hoffman, G. T. Mogel, R. Phillips, R. A. Robb, & D. Stredney (Eds.), *NextMed: Health Horizon* (pp. 127-130). Amsterdam: IOS Press.

Hubal, R. C., Fishbein, D. H., Sheppard, M. S, Paschall, M. J., Eldreth, D. L., & Hyde, C. T. (2008). How do varied populations interact with embodied conversational agents? Findings from inner-city adolescents and prisoners. *Computers in Human Behavior*, *24*(3), 1104-1138.

Kizakevich P. N., Hubal, R., Guinn, C., Starko, K., McCartney, M. L., & Magee, J. H. (2001, Summer). Virtual simulated patients for trauma and medical care (abstract). *Telemedicine Journal and e-Health*, *7*(2), 150.

Kizakevich, P. N., Duncan, S., Zimmer, J., Schwetzke, H., Jochem, W., McCartney, M. L., Starko, K., & Smith, T. (2004). Chemical agent simulator for emergency preparedness training. *Studies in Health Technology and Informatics*, *98*, 164-170.

Kizakevich, P. N., Furberg, R. D., Duncan, S., Hubal, R., Stanley, J., Merino, K., & Holloway, J. (2006). Mass casualty triage simulation for emergency preparedness and response [abstract]. *Telemedicine Journal and e-Health*, *12*(2), 165.

Krizowsky, P., Waters, H., Wright, M., Hubal, R., & Frank, G. (2008). *Dynamically configured scenarios for training adaptive network system operators*. Proceedings of the Interservice/Industry Training, Simulation and Education Conference (pp. 1392-1399). Arlington, VA: National Defense Industrial Association.

McMaster, L., Cooper, G., McLin, D., Field, D., Baumgart, R., & Frank, G. (2002). *Combining 2D and 3D virtual reality for improved learning*. Proceedings of the Interservice/Industry Training, Simulation and Education Conference (pp. 246-254). Arlington, VA: National Defense Industrial Association.

Nunez-Neto, B. (2008). *Border security: The role of the U.S. Border Patrol* (CRS Report to Congress, Order Code RL32562), Washington, DC: Congressional Research Service.

Robertson, I. T., & Downs, S. (1979). Learning and the prediction of performance: Development of trainability testing in the United Kingdom. *Journal of Applied Psychology, 64*, 42-50.

Robertson, I. T., & Downs, S. (1989). Work sample tests of trainability: A meta-analysis. *Journal of Applied Psychology*, *74*, 402-410.

Robertson, I. T, & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology, 55,* 171-183.

Siegel, A. I. (1983). The miniature job training and evaluation approach: Additional findings. *Personnel Psychology*, *36*, 41-56.

Siegel, A. I. (1986). Performance tests. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 121-142). Baltimore, MD: Johns Hopkins University Press.

Siegel, A. I., & Bergman, B. B. (1975). A job learning approach to performance prediction. *Personnel Psychology, 28*, 325-339.

Siegel, A. I., & Jensen, J. (1955). The development of a job sample trouble-shooting performance examination. *Journal of Applied Psychology, 39,* 343-347.

U.S. Department of Homeland Security (2009, February). *Independent review of the U.S. Customs and Border Protection's reporting of FY 2008 drug control performance summary report* (OIG-09-21). Washington, DC: Department of Homeland Security, Office of Inspector General.