DEVELOPMENT AND VALIDATION OF A COMPUTER INTERACTIVE TEST BATTERY

Valarie A. Sheppard, M.A.
Todd A. Baker, Ph.D
Deborah L. Gebhardt, Ph.D.
Kristine M. Leonard, M.A.

Human Performance Systems, Inc.
hpsvsheppard@erols.com

Paper presented at the 2000 IPMA-AC Conference, Arlington, VA

INTRODUCTION

The use of valid employment selection procedures for jobs with physical and perceptual motor demands has provided organizations with many benefits ranging from increased productivity to reductions in on-the-job injuries. Private sector organizations such as the freight, railroad, and electric industries and public sector agencies (e.g., U.S. Navy and Army, cities, states) have increased their use of selection assessments that evaluate these capabilities because of the benefits received. To ensure selection instruments meet fair employment standards and conform to federal and state guidelines, it is important that the essential job components be identified. Based on the analysis of the essential job components, physical and perceptual motor performance assessment procedures can be designed to ensure that new employees are qualified to perform a variety of jobs in a safe and efficient manner.

For reasons of safety and effective job performance, employers in the shipping industry must be concerned with selecting applicants for the Container Equipment Operator (CEO) position whose capabilities are commensurate with the demands of the job. Use of a fair and valid evaluation system helps the organization comply with the Americans with Disabilities Act (1990) and the Federal Statutes (e.g., Uniform Guidelines, Civil Rights Act of 1991) and assesses the applicants' ability to perform the essential functions of the job.

The purpose of this project was to develop valid evaluation procedures in the selection of Container Equipment Operators (CEOs). This report describes the development and validation of computer adaptive and paper and pencil tests that were predictive of the skills required for operation of heavy equipment.

JOB ANALYSIS

Identification of Essential Tasks

A job analysis was conducted to identify the essential tasks performed in the CEO job. The results of the job analysis will be used to develop the test battery for selection of CEOs.

Information such as job descriptions and job requirements for the CEO position were obtained and reviewed by Human Performance Systems, Inc., (HPS) staff. This information and previous job analysis results from the HPS database provided the framework for understanding the job and resulted in the development of the preliminary task list used during the site visit.

Site Visits

Site visits were conducted by HPS staff at several ports to observe the CEO job performance. HPS staff observed the operation of equipment (i.e., Crane, Transtainer, Top Loader, Straddle Carrier, Empty Handler, Stacker, Yard Hustler, Forklift) both from the ground and from inside each piece of equipment. Observation from the ground allowed staff members to view the entire task being performed (e.g., loading containers onto a railcar using a top loader, placing containers on a chassis using the crane). Staff members also viewed the internal operation of the equipment to experience the physical characteristics associated with the job. In addition to the observations, the preliminary task list was reviewed with a number of CEOs. These CEOs were asked to review the task statements for accuracy and clarity. Questions related to how a task was performed, the ergonomic parameters and the correct use of terminology were also addressed.

Ergonomic measurements such as the physical dimensions of equipment (e.g., height of steps, placement of controls, types of displays (e.g., digital, colored lights) and fields of vision were also obtained. These measurements were input into the appropriate tasks on the task list.

Following the completion of site visits, the preliminary task list was revised based on the information gathered during the site visits. This revised task list was used in the task inventory step of the job analysis.

Task Inventory

A task inventory was constructed to determine the essential tasks, and obtain ergonomic data related to the CEO position. The task inventory consisted of: (1) a background information sheet, (2) a list of tasks performed on each piece of equipment (e.g., transtainer), and (3) a list of supplemental questions. Tasks were grouped in categories such as climb, walk, operate equipment/drive, and vision for each piece of equipment. Each task was rated on four rating scales: Frequency, Importance, Time Spent, and Physical Effort. The Frequency rating identified the number of times the task was performed. Importance measured the criticality of the tasks to completion of the CEO job. Time Spent ratings measured how much time was spent performing the task, and Physical Effort was used to identify the physical demand of a task.

Raters were first asked to rate the frequency with which they operated each piece of equipment in the past year. If a single task was not performed in the past year or never performed, the raters were instructed to rate the task as a zero on the Frequency, Importance, Time Spent, and Physical Effort scales. For equipment that was rated as a one (1) or higher on the Frequency scale, the raters were instructed to complete the individual task ratings. These procedures ensured that ratings were supplied by individuals who were familiar with the piece of equipment and the tasks. A set of supplemental questions was also included to gather information about selected ergonomic parameters and working conditions present when CEOs perform selected job tasks. A total of 63 usable task inventories were collected from 63 CEOs with a mean age of 50.4 and 20.7 years as a CEO.

Determination of Essential Tasks

Descriptive statistics for each task were computed for each of the rating scales. The means and standard deviations for frequency were computed with a minimum value of zero (0) and with a minimum value of one (1). The minimum value of one (1) was used because it indicated the rater performed the task in the past year. The means for Importance, Time Spent, and Physical Effort were computed only for raters who rated Frequency as one (1) or greater.

A decision rule was developed to identify the essential job tasks for each piece of equipment. To be considered essential the task had to have a mean importance rating of 5.0 or higher, signifying that incumbents thought the task was "very important", "of great importance," or " extremely important" to successful job performance. Further, tasks with a mean importance rating of greater than or equal to 5.0 had to be performed by 25% or more of the rated for the specific piece of equipment. For the general tasks, which were rated by the entire sample, the task mean ratings of greater than or equal to 5.0 had to be rated by 25% or more of the total sample. A total of 86 essential tasks were identified for the CEO position.

Results of Supplemental Questions

The results of the supplemental questions indicated that CEOs walk approximately ¾ of a mile in an average work day. The average number of containers handled in an hour ranged from 7.5 containers for a Hustler to 25.7 for a Crane, and 11.3 for the Straddle Carrier, 17.5 for the Transtainer, and 22.2 for the Top Loader. In addition, Crane operators climbed the ladder to reach the cab more often (63.5%) than they took the elevator (36.5%).

Identification of Abilities

To develop a battery of tests for selecting CEOs the knowledge, skills and abilities required in the job must be identified. This requires the determination of the types of abilities (e.g., physical, perceptual) needed to operate equipment efficiently. To determine the specific physical requirements of tasks, the Physical Abilities Analysis (PAA) method was selected. The PAA method translates task characteristics into the abilities requirements needed, by linking tasks to specific physical abilities. This is accomplished by providing descriptive language that defines physical ability requirements as a function of task characteristics. This job analysis method produces a quantitative profile of the abilities judged to be required for each task.

The PAA methodology is particularly relevant to issues of content and construct validation, because the method provides the basis for demonstrating the job relevance of the ability tests selected and their linkage to essential job tasks. This procedure meets the need to link essential job tasks to objective assessment measures as defined by the EEOC Uniform Guidelines (1978). The identification of the specific physical abilities required for effective job performance was completed in two steps.

CEO Ability Requirements

In Step 1, six experienced trainers and 4 HPS job analysts used a 3 point scale to rate whether each ability was essential, helpful, or unnecessary for the performance of the essential job tasks. For each ability, a definition that explained the ability and examples of tasks that required the ability were provided. Each rater independently read each essential task and determined whether the ability was "essential, helpful, or not needed" for performance of a task.

Four to six raters rated the tasks on the 16 abilities. For these ratings, type 1 ICCs were calculated. The interrater reliabilities for the vision, hearing, dexterity, and physical abilities were high ranging from .69 to .89 for vision, .78 to .90 for hearing, .81 to .92 for dexterity, and .80 to .95 for physical. These high reliability coefficients indicate that the raters agreed on the necessity of each ability in the performance of the essential job tasks.

A decision rule was defined specifying 1.5 as the point at which the ability was necessary for the performance of the job task. A mean rating below 1.5 indicated the ability was helpful but not necessary, or that the ability was not necessary for the performance of that job task. A mean rating of 1.5 was selected because it is a conservative level that allows for inclusion of all abilities that define the job. Of the 16 abilities examined, the following 13 abilities were found to be important to the performance of various job tasks: *Near Vision, Far Vision, Peripheral Vision, Depth Perception, Hearing, Sound Localization, Reaction Time, Control Precision, Time Sharing, Dynamic Strength, Flexibility and Equilibrium.*

Determination of Ability Levels

Step 2 of this process involved the determination of the level of each ability required for the completion of the essential tasks. The determination of the overall magnitude of each ability enabled HPS to determine whether it was appropriate to evaluate the

ability in the CEO selection process. Three HPS job analysts used a revised task by abilities list containing the essential tasks with an ability rating of 1.5 or greater to independently rate the level of ability needed for that essential task. After the ratings were completed, a consensus meeting was held and a final decision was made on each task.

A mean rating of 3.0 or higher was selected as the level to warrant testing of that ability since it signifies that there is a moderate to high demand for that ability, while a mean rating below 3.0 indicated a lesser demand for the ability. A mean rating of 3.0 was also selected because it was a conservative level allowing for inclusion of all pertinent abilities. *Far vision, depth perception, hearing, dynamic strength, multi-limb coordination, reaction time, control precision,* and *time sharing* had mean ratings of 3.0 or higher. These abilities were addressed in the development of the selection system.

TEST DEVELOPMENT

The abilities necessary for success in the CEO position are divided into four categories: Vision Abilities, Hearing Abilities, Dexterity Abilities and Physical Abilities. Human Performance Systems, Inc developed a medical guidelines/evaluation system to assess the Vision, Hearing and Physical Abilities of the CEO position. However, it was determined that a computer interactive test battery should be used to assess a candidate's Multi-Limb Coordination, Reaction Time, Control Precision, and Time Sharing abilities.

Literature Review

Review of the literature indicated that several paper and pencil tests existed that were appropriate to assess the critical abilities (Tests in Print, 1994). It did not reveal however, any commercially available computerized tests for the abilities required to perform the CEO job. Instead, the literature review found that most of the computerized testing had been done for military organizations (e.g., U.S. Army's Project A study (Alderton, Wolfe, and Larson, 1997)). It was determined that obtaining approval to use these tests for the validation study would not only be time consuming, but would not guarantee approval for use of the tests in an ongoing selection battery. Therefore, it was decided that new computerized tests should be developed.

Paper and Pencil Test Selection

Based on the literature review, several paper and pencil tests were selected that were perceived to evaluate abilites similar to the computerized tests. The tests selected included: 1) Space Relations (Paper Puzzles) developed by Thurstone, 1951; 2) Perceptual Speed (Identical Forms) developed by Thurstone and Jeffrey (1938); and, 3) Visual Pursuit developed by G. Grimsley (1956).

Computer Test Construction

Three computerized tests were developed to assess multi-limb coordination, reaction time, control precision, and time sharing which were found to be critical to CEO task performance.

The tests were developed to run on a standard Pentium System computer with a SuperVGA monitor. The computer was also equipped with a Windows95 operating system, a sound card, two (2) serial joysticks, and a game port. Each joystick was configured to control a different range of movement of objects on the monitor screen. The right joystick controlled up and down movements, while the left joystick controlled left and right movements.

The three computerized tests developed for the CEO battery were the Maze Test, Maze and Ball Test, and Cargo Test.

Maze Test

The Maze test is a two-handed tracking task that measures control precision and multi-limb coordination. The task consists of a rectangular screen area with a single maze design. The target in the shape of a box appears at the beginning of the maze and moves along the maze path. The subject is required to position the crosshairs [+] on the box and maintain that position as the box moves along the path to the end of the maze. When the crosshairs are positioned correctly on the box, they change from white to black. Within each trial, the box moves at a constant speed through the maze. However, the speed of the box movement is different for each trial. The target speeds were 25, 31, 46 and 55 seconds.

Trial Speed and Order. Two factors had to be determined prior to finalizing the maze test. The first was the speed of box movement and the second was the order of presentation of the different speeds. Since this maze and the other computer tasks would be novel skills for applicants, the research literature was reviewed to determine the ranges for speed of movement and the ordering of the trials. Based on this research, target speeds of 10 seconds to 50 seconds were pretested on a small group of job analysts. Based on this pretest, pilot speeds of 20 seconds to 45 seconds were selected.

The order of the speeds has been shown to influence learning and performing new psychomotor skills. McCracken and Stelmach (1977) showed that people who practiced a new skill at a constant speed for all practice trials had less errors in the practice trials, but when they transferred to the test trials at a different speeds they had more errors. Individuals who practiced at all the speeds

used in the testing did better on a test trials than the constant group. Similarly, Shea and Morgan (1979) found that subjects who practiced at random speeds did better on test trials of random and blocked speed tests than subject who practiced at each speed in a block of trials.

The type of practice, mass versus distributed, is important to learning new skills. Past research has shown that distributed practice with multiple rest periods is better for learning a new skill than mass practice in which the subject does multiple trials with little rest (Adams and Reynolds, 1954; Stelmach, 1969). To optimize learning distributed trials were used in the practice phase.

Scoring the Maze Test. The scoring of this test includes: (1) time of target, and (2) average distance of the crosshairs from the target.

Maze and Ball Test

The Maze and Ball test is a two-handed tracking task that measures control precision, multi-limb coordination, reaction time, and time sharing. The Maze and Ball task consists of a maze identical to the one in the Maze test. However, in this test the individual must also respond to a distractor (red ball) that appears at different times during the test as the primary task of tracking the box through the maze track is performed. As in the Maze test, two joysticks are used to control the crosshairs.

The distractor task requires the individual to respond to a visual stimulus (a red ball) that appears and moves horizontally or vertically around the fixed border of the maze. The stimulus continues to move around the maze within the border area until the individual responds by pressing the space bar and making the ball disappear, or until the allotted time expires. Each time the ball appears in a different place. To ensure that the individual does not continually press the spacebar in anticipation of the appearance of the ball, the number of times the spacebar is depressed when the ball is not present is recorded and subtracted from their total score.

The individual must remove his/her hands from the joystick in order to press the spacebar. When the spacebar is depressed the box pauses its movement along the maze trial. This pause allows the individual to return his/her hands to the joystick and position the crosshairs over the box before it starts to move again. The target speed were 16, 22, and 25 seconds, with the stimulus (ball) flashing into the target area every 5 seconds.

Scoring the Maze and Ball Test. The scoring of this test includes the "time on target" and "average distance of the crosshairs from the target" that were recorded for the Maze test. A third score, "time to react to the stimulus," is assessed by timing the interval from the time the ball appears until the spacebar is depressed. The time to complete each movement when the ball appears during a trial is recorded. The times for each ball are summed to form a "total trial reaction time" score. A fourth score, "number of false hits," is assessed by counting the number of times the spacebar is depressed when the ball is not present on the screen (false positives).

Cargo Test

The Cargo test is a two-handed movement task that measures control precision and multi-limb coordination by moving boxes from one side of the screen to the other side. The Cargo test consists of a column of five (5) boxes on the right side of the screen and five empty places on the left side of the screen, with a barrier of non-movable boxes stacked in the middle of the screen. The goal of this task is to move the boxes (one at a time) from the right side to the left side of the screen as quickly as possible without touching the barrier in the middle. This task is accomplished by positioning the crosshairs over the box to be picked up and locking on, by pressing the joystick trigger. Once the box is locked it can then be moved to the other side of the screen, to an empty position in the left-hand stack. If while moving the box, the individual bumps the barrier in the center of the screen, the box is dropped and returned to its original position in the right-hand stack. The individual must then start again. The cargo test is self-paced with no time limit.

Scoring the Cargo Test. The scoring of this test includes: (1) number of boxes correctly moved during the allotted time for each trial; (2) total time to move all the boxes in each trial; (3) number of unsuccessful attempts to lock or drop a box; and, (4) number of times the individual bumps into an obstacle such as the middle stacks.

Test Instructions and Practice Trials

Test Instructions

All instructions for the test are presented through an audio prompt within the computer system. The instructions are presented orally since CEOs must respond to audio input via radios during the performance of their job duties.

The individuals are first presented with an overview of each test (e.g., Maze) that explains what is involved with a test, as well as the object of that particular test. They are then given detailed instructions for completing the test. In order to allow individuals to hear the instructions multiple times (e.g., if they do not understand what to do), a prompt appears at the end of the instructions asking whether they wish to hear them again or to continue with the test. If after repeating the instructions, the individual is still unsure of what is being asked for, an option is given for summoning the test administrator. The system then allows the test administrator to replay the instructions with his/her additional comments until the individual understands what is required.

When the individual understands the instructions, and indicates that he/she is ready to move on, a step-by-step walk through of the test is presented. During the walk through the system orally prompts them for each action (e.g., "Position the crosshairs over the box", "If the crosshairs are not black then you are not over the box"). After the walk through, the individual is given the opportunity to practice the test in a set-up identical to the test trials. When the practice trials are complete, the individual is prompted to get ready for the actual test.

The instruction procedures (e.g., demonstration, walk through) are identical for the Maze, Maze and Ball, and Cargo tests. In each test, the individual is given multiple opportunities to repeat the instructions and an opportunity to practice the test.

Computer Equipment Familiarity Training

In developing the three tests, it was recognized that individuals taking this test may be unfamiliar with the operation of a computer and/or the use of joysticks, especially the use of two joysticks, simultaneously. In order to give these individuals an opportunity to become more familiar with the operation of the equipment, a practice task was developed.

The practice task involved removing boxes from the screen by using the joy sticks to line up the crosshairs over the box, and then clicking on the joystick button to make the box disappear. The individual was presented with a series of boxes in different configurations of increasing difficulty. The individual first had to "delete" three boxes in a horizontal line, then three boxes in a vertical line, then six boxes scattered around the screen. Individuals were then presented with those three configurations a second time, but in this series of trials the boxes were moving.

In addition to allowing individuals to become familiar with the operation of the computer and the joysticks, these practice tasks also allow for familiarization with the auditory instructions that were used throughout the test.

Pretest of Computer Test Battery

Prior to using the three test computerized test battery in the validation study, the test battery was pretested to ensure that the speed of movement targets, the manipulation of the hardware (e.g., joysticks, spacebar), and scoring system functioned properly. The pretest was administered to sixty (60) subjects from a local university. The sixty pretest subjects consisted of 32 men and 28 women. Their ages ranged from 19 to 42 (mean = 21.8, S.D.=3.0). The sample was made up of 35 whites, 11 black, 2 Hispanics, 6 Asians and 6 missing. The subjects were each given six practice trials and 10 timed trials of each test in the test battery (i.e., Practice Task, Maze Test, Maze and Ball Test, Cargo Test). The trials for the practice task, the Maze test and the Maze and Ball test were presented at various speeds in a random order. At the end of the testing session the subjects completed the opinion questionnaire which provided feedback about the instructions and the operation of the system.

Based on information gathered during the pretest, several changes were made to the test battery. First, the instructions were revised in several places to improve the clarity of the sound and to make them easier to understand. Secondly, changes were made to the computer program to eliminate errors that became apparent during the pretest. Finally, the data from the practice trials were analyzed to determine the level of improvement in skill between trials. The analysis showed that improvement occurred during practice trials 1, 2, and 3 but did not occur during trials 4, 5, and 6. The analysis coupled with the feedback from the opinion questionnaire also indicated that there was an increase in fatigue by the third practice trial for the third test. Therefore, the number of practice trials was decreased, from six trials to three.

DEVELOPMENT OF JOB PERFORMANCE CRITERION MEASURES

To validate the computerized tests it was necessary to develop measures of job performance (i.e., criterion measures) that assessed the workers' ability to perform essential tasks. These measures were used to determine whether the tests were related to the workers' job performance. Three criterion measures were developed and used to determine the validity of the tests. These criterion measures were a work sample, self ratings, and observation and rating of actual equipment operation. The steps conducted to develop these criterion measures are described below.

Work Sample Criterion

The job analysis, supplemental questions, and PAA results were used to design a work sample that reflected the essential tasks associated with the operation of an entry-level piece of equipment. The intent of the work sample was to accurately reflect essential tasks performed. Typically, new CEOs are trained on the hustler prior to other pieces of equipment, and due to seniority they tend to operate the hustler for at least five years with only sporadic operation of the other equipment, a work sample consisting of operating a hustler was developed. This work sample included driving and making turns with and without a container, backing the hustler up and attaching to a container, and backing the hustler and container up into an 11-foot wide space. The work sample generated two scores, time to complete the course with split times taken at specified intervals and ratings of the individual's performance on each task.

A rating form was developed using incumbent input. For each task in the hustler course positive and negative elements were identified. A 5-point scale was selected with a "4" indicating superior/exceptional performance and a "0" indicating very poor performance.

Self Ratings

Since the CEOs perform most tasks by themselves, it was determined that peers and supervisors may have difficulty providing performance ratings because of inadequate opportunities to observe their co-workers. Therefore, a self-rating measure that assessed the essential tasks associated with different pieces of equipment was developed. Incumbents used this form to rate their performance on each piece of equipment that they operated in the past year. The self-rating form consisted of instructions for the raters, questions about the last time and how often they operated each piece of equipment, and a list of tasks to rate for each piece of equipment. The number of tasks for each piece of equipment ranged from two (empty handler) to six (straddle carrier). For each task on the self rating form, positive and negative elements were listed to help the individual make his or her ratings. The self-rating form also contained a brief set of questions about the type of equipment operated within the past year, and the frequency of operation.

Observation and Rating of Equipment Operation

A third criterion measure used to investigate the relationship between test performance and job performance was observation and rating of equipment operation by knowledgeable trainers and supervisors. For this criterion measure, an equipment operator trainer would observe the individual operating a piece of equipment (e.g., straddle, top loader) on the job and rate the individual's performance using a rating form. The ratings from the trainer served as the criterion measure. An equipment operation form for hustler was not developed since data on hustler operation performance was collected during validation testing.

As with the other criterion measures, incumbent input was used in the development of the ratings forms. This procedure was the same one used to develop the hustler and self rating forms. For each task and piece of equipment positive and negative elements were identified. The same 5-point scale used for the work sample was used for this criterion measure.

VALIDATION DATA COLLECTION

Validation of the tests (predictors) involves assessment of the relationship between the tests and job performance (criterion measures). A criterion-related validation strategy was selected for this study to demonstrate the statistical relationship between the computer and paper and pencil tests and job performance. These data were also used to establish a minimum cutoff score. A concurrent validation study was conducted in which incumbent CEOs ability were assessed using the computerized tests and paper and pencil tests. Participants' job performance was evaluated using the work sample, self ratings, and equipment ratings. Since the paper and pencil tests were standardized tests of abilities required for the CEO job, the relationship between scores on the paper and pencil tests and computer tests was examined to determine if the computer tests assess the abilities required for successful job performance.

Testing

Each CEO participated in a 4½ hour session. A total of seven sessions were conducted with 6-12 participants in each session. After the project was explained participants were placed in groups of 2-3 and assigned to one of three areas (hustler course, computer testing, paper and pencil testing and self ratings). Each group completed the three areas during the session. Those CEOs not qualified to operate the hustler did not complete the hustler course. All data for the three tests (maze, ma ze and ball, cargo tests) were gathered by the computer and placed into separate data collection files.

Criterion Measure Data Collection

Hustler course scores and ratings and self ratings were collected during the validation test session. Equipment ratings were completed on a date after the test sessions. Equipment raters were given a list of CEOs that they were to observe and rate. The raters scheduled a time to observe and rate these individuals operating a specified piece of equipment. The completed equipment rating forms were then sent to HPS.

A total of 54 of the job equipment ratings (i.e., 20 Crane, 22 Top Loader, 12 Straddle Carrier) were identified for completion. A total of 13 of these ratings were returned to HPS. The return rate for the usable equipment rating forms was 24.1%.

VALIDATION RESULTS AND DISCUSSION

Validation Sample Demographics

The validation sample consisted of 49 participants, 47 men and 2 women. The ages for the participants ranged from 23 to 63 years (Mean = 50.1, SD = 10.3). The ethnic breakdown of the sample included 18 Whites, 20 Blacks, 7 Hispanics, one other, and three missing.

The validation sample, in general, did not have experience with a computer keyboard or mouse. Approximately 84% of the sample (n = 41) had never used a computer keyboard or mouse. Of the 16% of the sample that had operated a computer, most used a keyboard or mouse only once a month. Ninety percent of the sample never operated the computer with one joystick, while 98% never used two joysticks. The percentage of the sample that played video games (29%) was higher than the percentages for keyboard, mouse, and joystick use. However, 12 of the 14 incumbents who stated that they play video games stated that they play only once a month. Thus, the sample of participants in the study did not have much experience operating computers, keyboards, mice, or joysticks. The experience operating computer equipment did not vary across ethnic groups.

Computer Test Results

Practice Trials Results

Analysis of the practice trials indicated that they helped participants improve their ability to operate the joysticks. Participants completed the second set of similar trials in faster times. For example, comparisons across trials 8 through 10 showed that the mean for trial 9 was 2.10 seconds faster than the mean for trial 8. Furthermore, the mean time to complete trial 10 was 1.02 seconds faster than the mean to complete trial 9.

Comparison of the two groups (video game experience, no video game experience) indicated that practice helped both groups improve their ability to operate the joysticks. The inexperienced group appeared to make larger improvements with practice than the experienced group. However, the inexperienced group was not able to obtain a similar level of performance as the experienced group.

Computer Test Performance - Maze Test

Four different speeds were used across the maze trials. These speeds were 55, 46, 31, and 25 seconds and reflected the time it took for the box to move through the maze. The box moved the slowest in the 55 second maze and fastest in the 25 second maze. Two scores were generated for each Maze test trial and are presented in Table 1. These scores were: (1) time on target (amount of time the participant kept the cursor on top of the moving box) and (2) average distance from target (the average distance the cursor was from the moving box for a trial). Comparisons among the trials with the same maze speeds indicated that participants improved their time on target and average distance from target scores in the later trials. The magnitude of the improvement in time on target scores was similar across the different maze speeds however, participants showed greater improvement in average distance from target scores at the higher speed mazes (speeds 31 and 25). This finding indicated that participants improved their scores by becoming more proficient in their ability to manipulate both joysticks simultaneously.

Comparisons of Maze test means across ethnic and age groups and video game experience indicated that there were no significant differences across ethnic groups, however there were significant age differences. The under 40 age group had better scores (greater time on target, lower average distance from target) than the 40 and over age group. Similarly, there were significant differences between the inexperienced and experienced test scores. For each of the 12 trials, the experienced group had better scores (greater time on target, lower average distance from target) than the inexperienced group.

Computer Test Performance - Cargo Test

Five scores were generated for each Cargo trial and are presented in Table 2. These scores were: (1) time to complete the task (move five boxes from one side of the computer screen to the other using two joysticks), (2) number extra locks (trying to lock onto a box when the cursor is not in the correct position), (3) number of extra drops (trying to place a box in the slot when the box is not in the proper position), (4) number of collisions with the stacks in the middle of the screen, and (5) lock time (time spent locked onto a box). Comparisons of time to complete means across the 10 trials indicated that participants improved their scores in the later trials, especially for trials 9 and 10. There was a difference of 7.56 seconds between the mean of trial 1 (85.32) and the mean of trial 10 (77.76). Improvements in the number of extra locks, extra drops, and collisions were not found across the 10 trials. However, the lock time means did decrease by 3.39 seconds from trial 1 (43.34) to trial 10 (39.95). This finding indicated that participants improved their scores by becoming more proficient in their ability to manipulate both joysticks simultaneously.

Comparisons of Cargo test means across ethnic groups, age and video game experience found significant differences on time to complete, lock time, and cargo final score means. The white group had better scores (faster scores) than the non-white group; the under 40 age group had better scores than the 40 and over age group and the experienced group had better scores than the inexperienced group.

Computer Test Performance - Maze and Ball Test

The mean score for Maze and Ball are presented in Table 3. Three different Maze and Ball speeds were used across the trials. These speeds were 25, 22, and 16 seconds and reflected the time for the box to move through the maze. The box moved the slowest in the 25 second maze and fastest in the 16 second maze. In addition, the ball appeared at 5 second intervals. The following three scores were generated for each Maze and Ball test: (1) time on target (amount of time the participant kept the crosshairs on top of the moving box), (2) average distance from target (the average distance the cursor was from the moving box for a trial), and (3) mean response time (mean response time to stimuli (ball) that appeared on the screen and depressing the spacebar to make the ball disappear). Comparisons among the trials with the same maze speeds indicated that participants did not show improvement in their time on target, average distance from target, and mean response time scores in the later trials. This lack of improvement may have been due to participant fatigue since this was the last of the three computer tests.

The comparisons of Maze and Ball test means across ethnic group, age group and video game experience found significant difference between the white and non-white groups (for all scores, the white group had better scores (greater time on target, lower average distance from target, faster mean response times) than the non-white group); between the under 40 and the 40 and over age group (the under 40 age group had better scores than the 40 and over age group) and between the inexperienced and experienced groups (the experienced group had better scores than the inexperienced group).

Computer Test Correlations

The correlations among the computer test scores from the Cargo, Maze, and Maze and Ball tests are presented in **Table 4**. Correlations included in **Table 4** were only those test scores that included all trials for a task. The following scores were included in the correlation analysis:

Maze Test: Mean time on target across all trials; Mean average distance from target across all trials; Grand Mean Maze final score across all trials

Cargo Test: Mean time to complete across all trials; Grand Mean Cargo final score across all trials

Maze and Ball Test: Mean time on target across all trials ; Mean average distance from target across all trials ; Mean average response time across all trials ; Grand Mean final score 1 across all trials ; Grand Mean final score 2 across all trials ; Grand Mean final score 3 across all trials

Correlations between test scores for the same computer test (e.g., Cargo task) were high. All correlations within a test, with the exception of mean response time for the Maze and Ball ranged from ±.81 to ±.99. Correlations between Maze and Ball mean response time and other Maze and Ball scores were moderate ranging from ±.60 to +.64. The reason for some of the negative correlations in Table 18 is that *lower scores* represent better task performance (i.e., Cargo time to complete, Cargo final score, Maze and Maze and Ball average distance from target, and Maze and Ball mean response time). For other scores such as Maze and Maze and Ball time on target and all Maze and Maze and Ball final scores, higher scores represent better task performance. Thus, a correlation of -.81 between Maze time on target and Maze average distance scores indicates that better performance on time on target (higher scores) is highly related to better performance on average distance (lower scores).

Correlations between the test scores from the three computer tasks were moderate to high. Correlations between the Maze and Ball mean response time score and the Cargo and Maze scores were lower ranging from -.50 to .48. With the exception of the Maze and Ball mean response time score, correlations between the Cargo and Maze scores and the Cargo and Maze and Ball scores were similar in magnitude. The correlations between Cargo and Maze scores ranged from ±.67 to ±.75. The correlations between Cargo and Maze and Ball scores ranged from ±.66 to ±.74. Correlations between Maze and Maze and Ball scores were higher ranging from ±.74 to +.94. The correlations between Maze and Maze and Ball were higher than the correlations between these tests and the Cargo task because of the similarity of the tasks. Maze and Maze and Ball were similar tasks (tracking a moving object), with the exception of the need to respond to stimuli during Maze and Ball trials. Thus, it was expected that both Maze test scores would be more highly related than the Cargo and both Maze scores.

Paper and Pencil and Computer Test Correlations

The correlations between the scores for the three computer tests and the three paper and pencil tests are **presented in Table 5** These correlations were examined to determine if the computer tests assessed abilities required CEOs. Overall, the correlations showed that the computer tests do assess Perceptual Speed and Visual Pursuit and, to some extent, Space Relations. All correlations between the Perceptual Speed test scores and computer test scores were significant at the $p < .01$ level. Correlations between Visual Pursuit test scores and all Maze test scores were significant at the $p < .01$ level. Other significant correlations ($p < .05$) for the Visual Pursuit test were with: both Cargo scores, Maze and Ball time on target, and Maze and Ball final score 2. The Space Relations test had significant correlations ($p < .05$) with three computer test scores: Maze time on target, Maze and Ball time on target, and Maze and Ball final score 2.

Criterion Measure Results

Three criterion or job performance measures are described in the criterion development section. These are: (1) hustler course performance (time to complete course and ratings of performance on the course); (2) self-ratings by subjects of their level of performance on multiple pieces of equipment; and, (3) supervisor/training personnel observation and evaluation of on-the-job performance on the straddle carrier, crane, top-loader, and empty handler. Although, adequate sample sizes were obtained for the hustler course and self-ratings, limited data were obtained for the equipment operation ratings. As a result, the equipment operation ratings were eliminated as a criterion measure. Due to the low number of subjects who operated the straddle carrier, crane, empty handler and top-loader, the self-ratings for these pieces of equipment were analyzed together. They were viewed from a standpoint of the most frequent pieces of equipment operated. This approach resulted in two combinations for the self-ratings: (1) mean z-score across all equipment, and (2) mean of the most frequently operated piece of equipment.

In addition to these two measures, two measures of the hustler course: (1) time to complete the hustler course, and (2) ratings of performance driving the hustler were included in the remainder of the analyses related to the criterion measures.

Criterion Measure Correlations

The correlation between the two self rating scores was high (.92). The correlation between the Hustler course mean rating and final time was low (-.33). This result indicated that the raters were assessing the participant's performance on the Hustler course using constructs such as accuracy and safety rather than speed. The correlations between the self rating and hustler course scores were low ranging from -.22 to -.09. These low correlations indicated that participant performance was being judged differently between the two measures. These low correlations suggested that both measures may have unique contributions to a combined measure of job performance (e.g., Hustler course mean rating + Self rating mean).

Formation of Overall Criterion Measures

Based on the correlation results of the hustler course and self ratings, seven criterion measure combinations were formulated. These combinations were made using the individual scores from the hustler course (final time, mean rating) and self ratings (mean rating, most frequent equipment mean rating). These criteria were: (1) self rating mean + hustler course final time, (2) self rating mean + hustler course mean rating, (3) self rating most frequent equipment mean rating + hustler course final time, (4) self rating most frequent equipment mean rating + hustler course mean rating, (5) hustler course final time + hustler course mean

rating, (6) self rating mean + hustler course final time + hustler course mean rating, and (7) self rating most frequent equipment mean rating + hustler course final time + hustler course mean rating.

Before combining self rating or hustler course scores, all measures were standardized (z-score). The standardized hustler course final time score was multiplied by -1.0 to reverse the direction of the scores. Before standardizing, higher hustler course final time scores reflected poorer performance. By multiplying the standardized score by -1.0, higher scores for the hustler course final time reflected better performance. This transformation made the hustler course time consistent with all other criterion measures (e.g., hustler course mean rating, self rating mean) and provided for easier interpretation.

Once all individual criterion measures were standardized, combinations were formed by summing different combinations of variables. The correlation between the hustler course time and mean rating was low indicating that the two measures were evaluating different aspect of performance. The correlation between the two self-ratings (mean across equipment and mean for frequent equipment) was high (r = .92). However, the self-ratings mean had a higher correlation with test performance. The goal was to predict performance on multiple pieces of equipment. Since the self-rating mean addresses multiple equipment and had higher correlations with the predictor tests, the self-rating mean was retained, and the self rating frequent mean was dropped from further analysis.

Selection of Final Criterion Measure

Based on these correlations, a final criterion measure was selected. Examination of the individual versus combined criterion measures found that most of the combined criterion measures had higher correlations with the tests than the individual criterion measures. Thus, the individual criterion measures were eliminated from consideration. Comparisons of criterion measures that included the self rating mean and the criterion measures that included the self rating most frequent equipment mean found that the self rating mean criterion measure combinations had higher correlations with the tests than the self rating most frequent equipment mean criterion measure combinations. Thus, the self rating most frequent equipment mean combination was eliminated from further analysis.

Since the hustler time to complete the course and hustler mean ratings assessed different aspects of the job, it was determined they be combined with the self rating mean. The self rating mean + hustler course final time + hustler course mean rating combination was selected as the final criterion measure. This combination had more significant correlations with the computer and paper and pencil tests than any other combination. Furthermore, the sample size for each criterion measure combination was similar (n = 29 to 31), so the loss of subjects resulting from the use of a particular criterion measure was not relevant. Therefore, the self rating mean + hustler course final time + hustler course mean rating criterion measure combination was used as the single criterion measure for all subsequent analyses.

Selection of Final Computer Test Scores

For each of the three computer tests multiple scores were generated. As shown in **Table 18**, the correlations among these scores within each test was high. Due to these high correlations, it was determined that only one score for each test would be used for the regression analyses. The final scores were selected because they all had significant correlations with the final criterion measure and they included multiple measures. For example, the Cargo final score included four measures (time to complete, number of extra locks, number of extra drops, and number of collisions). The computer scores used for subsequent analyses were: Cargo mean final score, Maze mean final score, and Maze and Ball final score 3.

Validity of computer and paper and pencil tests

Multiple Regression Analysis of Predictor Tests

A forward-backward stepwise multiple regression approach was used to examine the validity of the tests (Dixon, Brown, Engelman, Hill, & Jennrich, 1988). This approach resulted in a multiple correlation that indicated the degree of relationship between the independent variables (i.e., the computer and paper and pencil tests) and the dependent variable (i.e., final criterion measure of self ratings, hustler course ratings, and hustler course final time).

Initial Overall Regression. The subset of tests selected as predictors of the criterion measure were Cargo final score, Maze final score, Maze and Ball final score 3, Perceptual Speed, Space Relations, and Visual Pursuit. The names of the Cargo and Maze tests were shortened for ease of the reader.

Results of this regression found that two tests entered the regression equation significantly (Table 29). The Cargo final score entered first, followed by Visual Pursuit. The two tests resulted in a significant multiple R of .60 [$F(2,26)$ 3.59; $p < .05$] with the criterion measure. The squared multiple R of .36 indicated that 36% of the variance in the job performance (i.e., criterion measure) was accounted for by these two tests.

Regression Equation for the Test Battery

Using unstandardized beta weights (i.e., relative weights) the regression equations resulting from the regression analysis are:

Equation 1 - Regression Equation:

Criterion = 0.09620(Visual Pursuit) - 0.03326(Cargo Final Score) + 2.87838

Validity Analysis Across Subgroups

A differential prediction technique was used to determine whether the final test battery was fair across ethnic, age groups, and video game experience. This moderated multiple regression procedure provided a method to examine whether the separate regression equations for ethnic, and age subgroups differed significantly from the overall regression equation (Bartlett, Bobko, Mosier, & Hannan, 1978; Kerlinger & Pedhazur, 1973).

The test fairness analyses for the test battery across ethnic groups (White, Non-White) showed no significant intercept [$F(1,22) = 0.901$] or slope [$F(1,22) = 0.082$] differences. The test fairness analyses for age (under 40, 40 and over) also showed no significant intercept [$F(1,25) = 0.130$] or slope [$F(1,25) = 0.067$] differences. Finally, the test fairness analysis showed no significant intercept [$F(1,25) = 0.322$] or slope [$F(1,25) = 0.040$] differences to those individuals with and without video game experience. Therefore, the regression equations was fair across ethnic and age groups and to those individuals with and without video game experience.

Simplification of the Prediction Equation

The unstandardized beta weights from Equation 4 were simplified by multiplying each weight by 100, rounding the results to the nearest integer value, and eliminating the intercept values. This process simplified the equations for the purposes of calculating the combined final score from the regression equation. Furthermore, the signs for the Cargo Final Score and Visual Pursuit were reversed to eliminate negative values. The sign for the Cargo Final Score in Equation 4 was negative indicating a shorter time and less negative moves (e.g., collision with the middle stack). This sign was switched to a positive for Equation 6. The sign for the Visual Pursuit in Equation 4 was positive. This sign was switched to a negative for Equation 6. Although reversing the signs eliminated negative values, this resulted in lower combined score values representing better scores. The simplified equation for the test battery is shown below:

Equation 2 - Simplified Regression Weights:

Criterion = 3(Cargo Final Score) - 10(Visual Pursuit)

REFERENCES

Adams, J.A. & Reynolds, B. (1954). Effect of shift in distribution of practice conditions following interpolated rest. Journal of Experimental Psychology, 47, 32-36.

Alderton, D.L., Wolfe, J.H., & Larson, G.E. (1997). The ECAT battery. Military Psychology, 9(1), 5-37.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 31, 233-241.

Civil Rights Act of 1991, S. 1745, 102nd Congress, (1991).

Dixon, W. J., Brown, M. B., Engelman, L., Hill, M. A., & Jennrich, R. I. (1988). BMDP Statistical Software Manual. Los Angeles, CA: University of California Press.

Equal Employment Opportunity Commission. (1992). A Technical Assistance Manual on the Employment Provisions of the Americans with Disabilities Act. Washington, D.C.: Equal Employment Opportunity Commission.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978). Uniform Guidelines on Employee Selection Procedures. Washington, D.C.: Bureau of National Affairs, Inc.

Fleishman, E.A. (1954). Dimensional analysis of psychomotor abilities. Journal of Experimental Psychology, 48, 437-454.

Ghiselli, E. E. (1964). Theory of psychological measurement. New York: McGraw-Hill.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York, NY: Holt, Rinehart & Winston, Inc.

McCraken, H.D. & Stemach, G.E. (1977). A test of the schema theory of discrete motor learning. Journal of Motor Behavior, 9, 193-201.

Murphy, L.L., Close Conoley, J., & Impara, J.C. (Eds.). (1994). Tests in Print IV. Lincoln, Nebraska. University of Nebraska Press.

Shea, J.B. & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. Journal of Experimental Psychology: Human Learning and Memory, 5, 179-187.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420 - 428.

Stelmach, G.E. (1969) Efficiency of motor learning as a function of intertrial rest. Research Quarterly, 40, 198-202.

Figure 1
Container Equipment Operator Abilities

Vision Abilities

1. Near Vision: The ability to see close environmental surroundings.

   Read the fine print of legal journals; Plug in a television set

2. Far Vision: The ability to see details of objects at a distance.

   Determine make and model of a vehicle 100 yards away.

3. Peripheral Vision: The ability to see and perceive objects located to the sides of one's field of vision.

   Monitor child sitting in passenger seat while driving vehicle.

4. Depth Perception: The ability to judges the distance of one or more objects from the observer and the distance of several objects from each other.

   Determine which of two distant buildings is closer.

Hearing Abilities

5. Hearing: The ability to detect & discriminate sounds that vary over a broad range of pitch &/or loudness.

   Identify that someone is calling out your name.

6. Sound Localization: The ability to identify the direction from which a sound originated.

   Locate someone calling out your name in a crowd; Identify the location of a noise.

7. Oral Expression: The ability to use English words or sentences in speaking so others will understand.

   Give directions to a motorist so that he can reach his destination.

Dexterity Abilities

8. Multi-Limb Coordination: The ability to use two or more limbs at the same time (e.g., both arms).

   Drive a vehicle with manual transmission.

9. Reaction Time: The speed with which a single response can be made when a signal or event requires it.

   Hit a baseball; Respond to the brake lights from the car directly in front of you.

10. Control Precision: The ability to make very accurate movements of the arms or legs to exact locations.

    Cut out an outlined pattern using a jigsaw.

11. Time Sharing: The ability to shift back & forth efficiently betwn 2 or more activities & sources of info.

    Watch street signs while driving vehicle.

12. Multi-Control Manipulation: The ability to manipulate two or more controls at the same time.

    Play a computer game

Physical Abilities

13. Static Strength: The ability to use muscle force for a single task lasting less than one minute which involves lifting, pushing, pulling, or holding objects.

    Lift a 50 pound box and carry it 20 feet; Push a loaded hand truck 100 feet.

14. Dynamic Strength: The ability to use muscle force for single, multiple tasks involving lifting, pushing, pulling, holding, or carrying which lasts two minutes or more. This ability also involves supporting or moving one's own body weight. It represents muscular endurance and emphasizes the resistance of muscles to fatigue.

    Stack fire wood in a pile for 10 minutes; Shovel snow out of a driveway for 15 minutes.

15. Flexibility: The ability to bend, stretch, twist, or reach out with the body, arms, or legs.

    Bend to look under furniture for car keys; Stretch to get object from top shelf of cabinet.

16. Equilibrium: The ability to keep or regain one's balance, or to stay upright when in an unstable position. This ability includes being able to maintain one's balance when changing direction while moving or when standing motionless.

    Paint ceiling of room while standing on ladder; Walk across an icy parking lot.

Table 1
Mean Scores on the Maze Test Across Trials

Trial	Maze Time	Time on Target M (SD)	Avg Dist from Target M (SD)	N
Trial 4	46	34.64 (18.5)	23.73 (30.2)	48
Trial 5	55	36.79 (18.4)	20.19 (25.9)	48
Trial 6	31	28.58 (16.9)	34.08 (34.8)	48
Trial 7	46	36.17 (19.7)	21.65 (31.6)	48
Trial 8	55	41.02 (19.8)	19.98 (34.9)	48
Trial 9	31	29.50 (17.1)	35.92 (44.7)	48
Trial 10	25	21.56 (12.4)	41.04 (44.2)	48
Trial 11	46	37.17 (19.5)	22.10 (30.1)	48
Trial 12	31	32.72 (17.4)	26.34 (31.8)	47
Trial 13	25	23.68 (15.4)	37.96 (44.0)	47

Table 2
Mean Scores on the Cargo Test Across Ethnic Groups

Score	Units	Total Score M (SD)	White M (SD)	Non-White M (SD)	t-value
Time to Complete Mean Across Trials	secs	81.04(22.0) n=47	70.47(16.0) n=17	88.31(24.3) n=25	-2.65*
Extra Locks Mean Across Trials	#	3.58(3.9) n=47	3.01(2.6) n=17	3.91(4.9) n=25	-0.69
Extra Drops Mean Across Trials	#	3.05(4.9) n=47	1.77(1.5) n=17	3.98(6.4) n=25	-1.39
Collisions Mean Across Trials	#	0.55(1.0) n=47	0.70(1.1) n=17	0.50(1.0) n=25	0.61
Lock Time Mean Across Trials	secs	41.6(11.3-) n=47	35.93(7.0) n=17	45.43(12.9) n=25	-2.78**
Grand Mean Final Cargo Score		83.41(23.1) n=47	72.27(16.4) n=17	91.08(25.7) n=25	-2.66*

* p < .05 ** p < .01

Table 3
Means Scores on the Maze and Ball Test Across Trials

Trial	Maze Time M (SD)	Number Stimuli M (SD)	Time on Target M (SD)	Avg Dist from Target M (SD)	Avg Response Time M (SD)
Trial 4	22	5	39.30 (19.4)	18.04 (24.6)	1.28 (0.4)
Trial 5	25	6	42.02 (18.0)	13.13 (13.8)	1.23 (0.3)
Trial 6	16	4	30.76 (17.7)	23.98 (26.9)	1.27 (0.4)
Trial 7	22	5	35.59 (18.2)	20.15 (28.0)	1.29 (0.4)
Trial 8	25	6	41.26 (18.9)	16.63 (26.5)	1.25 (0.3)
Trial 9	16	4	33.15 (19.5)	22.30 (31.0)	1.30 (0.4)
Trial 10	25	6	40.07 (18.8)	16.61 (21.1)	1.36 (0.3)
Trial 11	22	5	36.82 (18.8)	18.67 (21.5)	1.32 (0.4)
Trial 12	16	4	32.60 (18.8)	23.09 (23.7)	1.38 (0.4)
Trial 13	22	5	38.02 (18.5)	18.04 (23.3)	1.33 (0.4)

N = 46 for trials 1 through 7 N = 45 for trials 8 through 10

Table 5
Correlations Between Computer and Paper & Pencil Test Scores

Computer Test Score	Perceptual Speed	Space Relations	Visual Pursuit
Cargo Time to Complete	-.57	-.21	-.30
Cargo Grand Mean Final Score	-.57	-.22	-.33
Maze Time on Target	.60	.30	.39
Maze Average Distance	-.47	-.18	-.42
Maze Grand Mean Final Score	.54	.21	.44
Maze and Ball Time on Target	.63	.33	.35
Maze and Ball Average Distance	-.39	-.20	-.21
Maze and Ball Response Time	-.42	-.19	-.26
Maze and Ball GM Final Score 1	.53	.28	.28
Maze and Ball GM Final Score 2	.63	.34	.35
Maze and Ball GM Final Score 3	.53	.28	.29

n = 46 p < .05 = .291 p < .01 = .376