AD-A054 953 HUMAN RESOURCES RESEARCH ORGANIZATION ALEXANDRIA VA F/G 5/9
MILITARY TESTING: KNOWLEDGE AND SKILLS; CRITICALITY AND RELIABI--ETC(U)
FEB 78 W OSBORN, J P FORD, C H CAMPBELL

UNCLASSIFIED HUMRRO-PP-4-78 NL

| OF |
AD
A054953

END
DATE
FILMED
7-78
DDC

(14) HUMRRO-PP-4-78

HumRRO

Professional Paper 4-78

HumRRO-PP-4-78

(9) Professional paper,

AD A054953

AD No. ____
DDC FILE COPY

(6) Military Testing: Knowledge and Skills; Criticality and Reliability.

(2)

(16) William/Osborn, J. Patrick/Ford, Charlotte H/Campbell, Roy C./Campbell James H./Harris

Four papers presented at the 19th Conference of the Military Testing Association San Antonio, Texas October 1977

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

(11) Feb[illegible]78

(12) 34p.

78 06 12 141

405260

PREFATORY NOTE

This paper is based on four presentations given at the 19th Conference of the Military Testing Association by HumRRO research scientists located at Fort Knox and Louisville, Ky.

The Osborn/Ford paper is based on work accomplished in Project SYNTEST, "Research on Methods of Synthetic Performance Testing." The Campbell/Ford/Campbell paper is based on work currently in progress in Project VALID, "Development and Evaluation of Self-Instructional Materials on Validation Procedures for Skill Qualification Tests."

The paper by Harris, Osborn, and Boldovici also is based on a currently active research effort, Project KNOX, "Tank Systems Skills and Training Structure." The final paper by Boldovici, Osborn, and Harris comes from work accomplished in Project PRETAC, "Derivation of Performance Requirements for Small Armor Units."

The Military Testing Association conference was held at San Antonio, Texas, October 17-21, 1977, with the Air Force Human Resources Laboratory and the Air Force Occupational Measurement as co-hosts.

TABLE OF CONTENTS

KNOWLEDGE TESTS OF MANUAL TASK PROCEDURES

William Osborn and Patrick Ford
HUMAN RESOURCES RESEARCH ORGANIZATION

Paper for:

Military Testing Association Conference
San Antonio, Texas

October 1977

KNOWLEDGE TESTS OF MANUAL TASK PROCEDURES[1]

The high cost of hands-on performance testing tends to complicate life for the developer of job proficiency tests. He is urged by reasons of economy to develop tests that are administratively feasible. This usually means tests that can be administered on a group basis—an interpretation that invariably leads to paper-and-pencil knowledge testing.

We know that knowledge tests are appropriate for tasks that are essentially mental, and we know they are inappropriate for tasks that involve finely tuned motor skill. But what of job tasks in between—tasks that involve both manual and mental activity? Many job tasks appear to be predominantly manual, but not particularly skilled. Placing some machines in operation, assembling objects, installing or repairing components, represent tasks that are essentially manual, but which, if performed without rigid time limits, cannot be considered psychomotor skills. This is not to say that such tasks require no skill. They must be learned, and if one identifies the skilledness of a task generally in terms of the amount of practice required to become proficient, then the aforementioned tasks are to some degree skilled. But the skilled aspect is probably mental, since knowledge must be acquired of what steps to perform, in what order and with what result. It may be hypothesized, in fact, that such manual task procedures can be performed with little or no practice, <u>if</u> one <u>knows</u> what, when and how to perform them.

If there is something to this hypothesis, proficiency can be measured validly in a knowledge testing mode, given one additional assumption: that the test medium is relatively neutral with respect to examinee differences in mental ability. This second assumption is necessary because we are considering a medium for testing that has no relevance to the medium for task performance. In other words, we would expect someone who can perform a task to be able to pass a hands-on test of that task; but if that person can't read or write at all well, we would be dubious of their ability to read and interpret written questions <u>about</u> task performance. It seems important, therefore, when substituting for a hands-on test, that the substitute medium not favor one type of examinee over another. We should strive to use test media that are neutral with respect to task-irrelevant differences in abilities.

With this perspective, I would like to describe an experiment in which we evaluated the validity of knowledge tests as substitutes for hands-on tests of manual task procedures.

The experiment was designed to examine four methods of knowledge testing in terms of their relative and absolute correlation with hands-on task proficiency for high and low mental ability subjects (*Ss*). The specific research questions of interest were:

- Do the four types of knowledge test correlate with hands-on task mastery?
- Do the types of test differ with respect to how well they distinguish masters from nonmasters?
- Do the types of test distinguish masters from nonmasters equally well for high and low mental ability levels?
- Do the types of test tend to produce the same <u>kinds</u> of errors in predicting task mastery?

[1] This paper is based on research done under Contract No. DAHC 19-74-C-0059 with the U.S. Army Research Institute for the Behavioral and Social Sciences. Conclusions and opinions expressed are the authors', and not necessarily those of the U.S. Army.

Method

Test Development. Tests were developed for three Army tasks: Installation of the Field Telephone (TEL), Setting up a Mechanical Ambush with the Claymore (AMB), and Disassembling the M-16 Rifle (RIF). The first two are clearly low-skilled tasks. Rifle disassembly, however, would be classified more accurately as moderately skilled, since some of the steps entail manipulations that are not easily mastered in one or two trials. Each task was analyzed into steps on which the test items were based. In addition to a performance (hands-on) test, four versions of a knowledge test were developed for each task. One version was a conventional multiple-choice test. The other three employed pictures in an effort to minimize literacy demands, but used different methods of eliciting task knowledge. A description of the four tests follows.

- Written Choice (WC). This is a standard multiple-choice test consisting of one question for each step in the task. A question focused on recognition of how a step is performed, when it is performed, or what its correct outcome is. Alternative answers to a question were limited to realistic options; unrealistic distractors were avoided. The test was scored by giving one point for each correct answer; seven was the maximum possible score for the TEL and AMB tasks, and eight the maximum for RIF.
- Picture Choice (PC). This method included the same questions as the Written Choice, but photographs were used in place of the printed word in presenting answer alternatives. The possible points and scoring procedure were the same as for WC.
- Picture Outcome (PO). In this method a photograph of the result of an improperly performed task was presented. Ss were instructed to inspect the picture and circle any errors. This type of test focuses on recognition of correct task outcome only. Test score was based on one point for each error circled, minus one point for each non-error circled. Total score was not allowed to go below zero. The possible range of scores was from 0 to 4 for TEL and RIF, and 0 to 3 for AMB.
- Picture Sort (PS). Photographs of steps in task performance, including both correctly and incorrectly executed steps, were used in this test method. The pictures were scrambled and presented to S with instructions to select the correct steps and place them in the order they should be performed. This method was considered to be the most comprehensive in its coverage of task knowledge; what steps to perform, and how and when to perform them are required knowledge. The method relies on recognition, as do the others, but all task elements are tapped and the guessing factor is minimized. Scoring was based on the award of one point for each picture or group of pictures representing a correct step performed in proper sequence. If two correct steps were in improper order, credit was withheld for the first step. Steps were judged to be improperly sequenced only if it were impossible or hazardous to perform them in that order. Maximum possible score was seven for TEL, and eight for AMB and RIF.

Subjects. Thirty-seven soldiers from units at Fort Knox were tested. They were chiefly from combat arms MOSs and ranged in grade from E-2 to E-6. For the purpose of study design, *S*s were in two mental ability (MA) groups: GT over 110 (high MA), and GT under 90 (low MA).[1] Twenty *S*s were in the high MA group and 17 were in the low.

Procedure. On arrival at the test site the project was explained briefly to *S*s. What was said to them took the following general form:

> "We are working on a project to evaluate several different methods of testing. You will take a hands-on test for three tasks. Then you will take four other kinds of tests for each task. After the test we will ask your opinion of it. This is not an MOS test, so there is no reason for you to be nervous. But the project is very important so, of course, we expect you to do as well as you can on every test."

All testing was done individually and began with administration of the hands-on test. At this point some *S*s received training on the task before going on to the knowledge tests. This was done to control the range of task mastery within the two MA groups. The intention was to create a rectangular distribution of mastery, with approximately a third of each MA group being wholly unqualified on a task, a third being partially qualified, and a third full masters. This approach worked well at the full mastery level since only one *S* could perform a task (TEL) without further training. Thus, 7 masters were created in each MA group by training them to pass the three hands-on tests. The approach did not work as well within the nonmastery range since most *S*s could perform some steps in the TEL and RIF tasks; only with the AMB task were any Ss trained to partial mastery.

Once an *S* had completed the hands-on test for a task, he was given the four knowledge tests successively. The order of test administration was counterbalanced over *S*s.

In addition to test performance, *S*s were asked their opinions of the methods by having them rank them from 1 to 5 with respect to the question: "Do you think this test is a good way to find out if a soldier can (task statement)?"

Scores on the 15 tests—one hands-on and four knowledge tests for each of three tasks—and *S*s ratings comprised the data that were analyzed.

Results

Continuous score correlations between knowledge test and hands-on performance for the three tasks are shown in Table 1 for the two levels of mental ability and for the total sample. With few exceptions the correlations are both statistically and practically significant. They are uniformly higher, regardless of test method, for the TEL and AMB tasks than for RIF, indicating that rifle disassembly is somehow different from the other tasks; a difference attributable perhaps to a more skilled motor component.

Comparison by type of knowledge test, for the total sample and total performance on the three tasks, indicates that the Written Choice, Picture Choice and Picture Outcome correlate equally well (.83, .80, and .84 respectively) with hands-on performance. The Picture Sort method yields a somewhat smaller overall relationship (.58), although the reduction is attributable to the near-zero correlation for the RIF task. The trend

[1] The GT (General-Technical) is a combination of scores on a verbal and a quantitative aptitude test. It is considered to be the best indicator of general mental ability in the Army Classification Test Battery.

Table 1

Correlations Between Performance and Knowledge Test Method for High and Low Mental Ability Groups

Mental Ability Group	N		Knowledge Test Method: Written Choice Task[a] TEL	AMB	RIF	TOT	Picture Choice TEL	AMB	RIF	TOT	Picture Outcome TEL	AMB	RIF	TOT	Picture Sort TEL	AMB	RIF	TOT
High	20	r	.69	.55	.17	.79	.71	.66	.47	.82	.76	.77	.31	.78	.70	.62	.31	.52
Low	17	r	.73	.82	.75	.90	.80	.76	.51	.80	.68	.74	.65	.90	.79	.56	.29	.69
TOTAL	37	r	.71	.67	.49	.83	.75	.70	.51	.80	.72	.74	.55	.84	.72	.55	.04	.58

[a]TEL = Installing Field Telephone
AMB = Installing Mechanical Ambush with Claymore Mine
RIF = Disassembling M16 Rifle
TOT = Total Performance on the Three Tasks

toward higher correlations for total score than for task scores reflects a tendency for intercorrelations among tasks to be lower for a knowledge test than for the hands-on criterion.[1]

Further analyses of the effectiveness of the different knowledge tests to distinguish masters from nonmasters, both within and between levels of mental ability, were carried out by analysis-of-variance. This is a reasonable way to examine the data, since mastery level was more of a manipulated "treatment" effect than a natural variate. Knowledge test performance, summed over tasks, of masters and nonmasters by mental ability level is shown in Table 2. All test methods did not have the same scale of measurement, so an ANOV (Winer, 1962) was performed on each method. Results of the four unweighted means ANOV are summarized in Table 3 and shown graphically in Figure 1. A clear and substantial main effect is revealed for mastery level, which merely represents the high correlations between knowledge test and task performance already mentioned. The size of this main effect for Picture Outcome relative to other test methods is worthy of note. The graphs in Figure 1 indicate that masters tend to average about five points higher than nonmasters on all tests, even though the potential range of performance on PO is only half that of the other tests. This would imply that a longer test would produce greater improvement in discrimination between masters and nonmasters for PO than for the other methods.

Performance on the knowledge tests tended to be lower for low mental ability *Ss* than for high, as indicated by the slope of the curves in Figure 1. The difference is small, and in fact not statistically reliable according to the separate ANOVs. However, when performance was converted to standard scores within test method and aggregated over methods, the mental ability factor is marginally significant ($p > .05$). Moreover,

[1]The reader will recall that, by design, the same people were masters on all tasks (had maximum criterion scores) although nonmasters varied in degree of nonmastery from task to task.

Table 2

Knowledge Test Performance (Means and Standard Deviations) of Masters and Nonmasters by Test Method and Mental Ability Level

Mastery Level	Mental Ability		Test Method: Written Choice	Picture Choice	Picture Outcome	Picture Sort
Masters	High	$\bar{X}$	18.71	20.28	10.29	18.71
		s	2.98	1.60	.76	3.25
		N	7	7	7	7
	Low	$\bar{X}$	17.86	19.29	10.14	17.57
		s	1.95	2.10	1.21	3.41
		N	7	7	7	7
Nonmasters	High	$\bar{X}$	13.69	15.00	6.38	15.15
		s	2.56	2.80	1.61	4.56
		N	13	13	13	13
	Low	$\bar{X}$	11.30	13.60	5.00	12.00
		s	1.83	2.59	1.41	3.06
		N	10	10	10	10

Table 3

ANOV Summaries of the Effects of Task Mastery (*M*) and Mental Ability (*A*) on Knowledge Test Performance

Test Method	Source	SS	df	MS	*F*
Written Choice	*M*	289.853	1	289.853	46.10[a]
	A	22.691	1	22.691	3.61
	M x A	5.126	1	5.126	.82
	Error	207.466	33	6.287	
Picture Choice	*M*	260.120	1	260.120	43.73[a]
	A	12.357	1	12.357	2.08
	M x A	.364	1	.364	.06
	Error	196.273	33	5.948	
Picture Outcome	*M*	177.034	1	177.034	95.38[a]
	A	5.060	1	5.060	2.73
	M x A	3.271	1	3.271	1.76
	Error	61.2483	33	1.856	
Picture Sort	*M*	180.178	1	180.178	12.73[a]
	A	39.781	1	39.781	2.81
	M x A	8.733	1	8.733	.62
	Error	466.939	33	14.150	

[a] $p < .01$

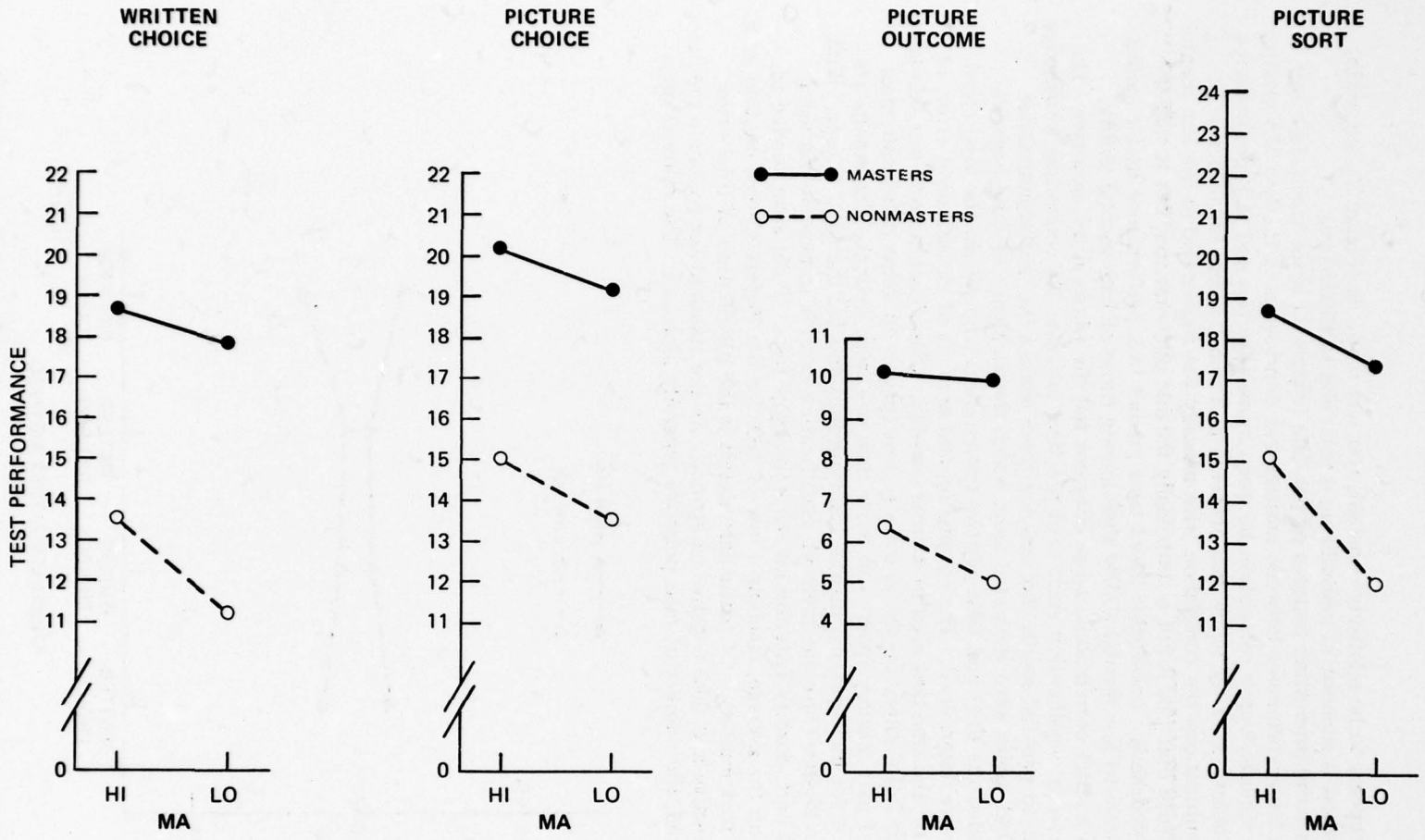

Figure 1. **Mean Performance of Masters and Nonmasters by Mental Ability (MA) Level for the Four Knowledge Test Methods**

the difference appears to be relatively constant over test methods (Figure 2), suggesting that no one method is superior in neutralizing mental ability differences.

One of the more interesting features of the data (Figure 1) is the trend, however slight, toward a larger difference between masters and nonmasters in the low MA group. This indicates a slightly higher correlation between knowledge test and task performance for low mental ability *S*s, a tendency also observed in Table 1 where for 9 of the 12 method/task combinations the correlation with mastery was higher within the low mental ability group. Note that this is not a statistically reliable phenomenon, but it suggests an interesting hypothesis: knowledge based tests predict task performance better among people of moderate to low mental ability than among those of high mental ability.

Validity in a strict correlational sense does not tell the whole story, however. The type of prediction or classification error is of practical interest. By converting knowledge test performance to pass-fail scores and arraying them against the master-nonmaster criterion, four-fold tables were generated from which the incidence of false negative (masters who failed the test) and false-positive (nonmasters who passed the test) classification errors were determined. The correlation and amount of classification error, of course, depend on the standard used in scoring pass-fail. Classification error was tabulated for a standard of full mastery on the knowledge test (pass = all items right) and again for a standard of part mastery (pass = no more than one item wrong). The results are shown in Table 4 for high and low mental ability groups and for the total sample. With exception of the Picture Outcome method, classification error is somewhat less using the more liberal part mastery criterion on the knowledge tests. Total error tended to run about 25% on the average, reaching a low of 16% for the Picture Choice method with the criterion of part mastery. Of particular interest is the distribution of total error between false-positive and false-negative categories. As the standard for passing a predictor measure is relaxed, the number of false-positives generally increases. The optimal ratio

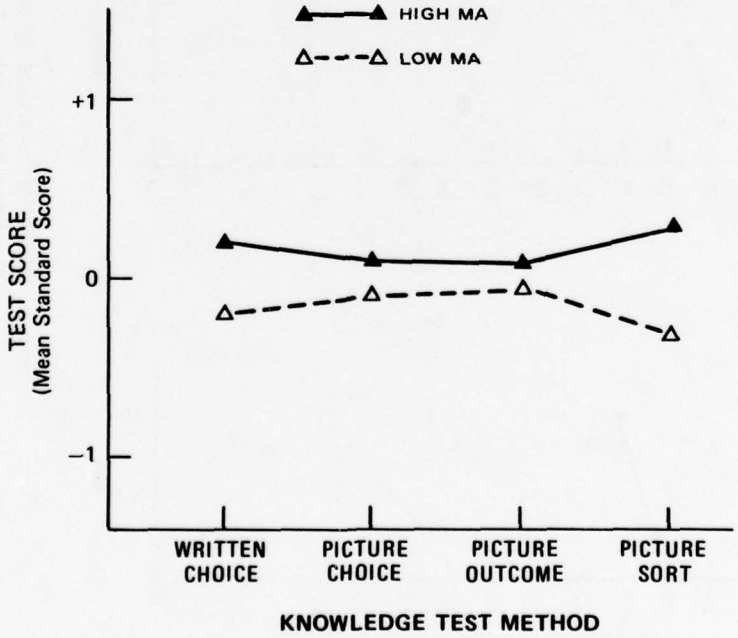

Figure 2. **Mean Standard Score Performance of High and Low Mental Ability (MA) Groups for the Four Knowledge Test Methods**

Table 4

Average[a] Percent Classification Error as a Function of Knowledge Test Method and Level of Mental Ability

Test Standard	Menta. Ability Group	Knowledge Test Method											
		Written Choice			Picture Choice			Picture Outcome			Picture Sort		
		Classification Error[b]											
		FN	FP	TOT	FN	FP	TOT	FN	FP	TOT	FN	FP	TOT
Full Mastery	High	18	05	23	13	07	20	05	15	20	23	03	26
	Low	27	02	29	25	04	29	16	08	24	33	00	33
	TOTAL	22	04	26	19	05	24	10	12	22	28	02	30
Part Mastery	High	07	13	20	02	13	15	00	32	32	15	17	32
	Low	18	02	20	08	10	18	04	20	24	22	02	24
	TOTAL	12	08	20	04	12	16	02	26	28	18	10	28

[a]Averaged over the three tasks.

[b]FN = False Negatives (masters who failed knowledge test).
FP = False Positives (nonmasters who passed knowledge test).
TOT = Total Classification Error.

of the two types of error is a moot point, and will depend largely on how test scores are to be used. But if test fairness is the goal, then minimizing the number of false-negatives should be the objective. The relative number of false-negatives, moreover, should be the same for groups differing in mental ability (or any other ability correlated with test score but unrelated to criterion performance). Comparing high and low MA groups we find a small but consistent tendency toward more false-positives among the high MA's, and more false-negatives among the low. This trend was evaluated by Chi-square analysis of the difference in type of classification error between high and low MA groups, and is shown in Table 5 by test method for each standard of test "mastery." Observed Chi-squares

Table 5

Chi Square of the Difference in Type of Classification Error Between High and Low Mental Ability Groups by Test Standard and Test Method

Test Standard	Knowledge Test Method			
	Written Choice	Picture Choice	Picture Outcome	Picture Sort
Full Mastery	1.33	1.54	4.19[a]	2.26
Part Mastery	7.22[a]	2.49	3.38[a]	6.30[a]

[a]$p < .10$.

were tested at the 10% level of significance, which provides for a conservative decision with respect to accepting the null hypothesis of no difference between groups in distribution of classification error. Type of classification error produced by the knowledge tests does appear to interact with mental ability. Although the number of cases underlying the analysis are too few to warrant firm conclusion, indications are that if one were interested in minimizing the incidence of false-negatives (i.e., the part mastery standard), the Picture Choice method produces the most equitable results for both mental ability groups.

Personal Preferences for Test Methods. *Ss*' opinions of the test methods were solicited after each test was administered and again when all testing was concluded. Responses at the two points in time were similar, so only the final ratings are reported here. *Ss* were asked to rank the five methods (including the hands-on criterion test) from highest to lowest in terms of the question, "Do you think this test is a good way to find out if a soldier can . . (e.g., set up a mechanical ambush with a Claymore?") Rankings were done separately for each task. Overall mean preference was highest for the hands-on method of performance testing, as might be expected (Tables 6 and 7).

Table 6

Mean Order of Preference[a] by Task for The Hands-On and Knowledge Test Methods

Task	Test Method				
	Hands-On	Written Choice	Picture Choice	Picture Outcome	Picture Sort
TEL	1.14	3.92	3.03	3.58	3.33
AMB	1.25	3.78	2.94	3.72	3.31
RIF	1.08	3.67	3.14	3.47	3.64

[a]The lower the number the higher the preference.

Table 7

Mean Order of Preference by Subgroup for the Hands-On and Knowledge Test Methods

Subgroup	Test Method				
	Hands-On	Written Choice	Picture Choice	Picture Outcome	Picture Sort
Masters	1.00	3.85	3.08	3.38	3.69
Non-Masters	1.30	3.83	2.65	3.87	3.35
High Masters	1.35	4.10	2.65	3.70	3.20
Low Masters	1.06	3.50	3.00	3.69	3.81
TOTAL	1.19	3.83	2.81	3.69	3.47

Differences in preference for the four methods of knowledge testing were less pronounced, although the Picture Choice consistently received higher average ranking regardless of the referent task or rating subgroup. Overall, the hands-on method was first, Picture Choice second, Picture Sort third, Picture Outcome fourth, and Written Choice last in average order of preference.

Discussion

A number of interesting though tentative findings emerged from this study. The small sample of people and tasks certainly limits generality of the results, and the following interpretation and conclusions should be so tempered.

The data strongly support the hypothesis that performance on manual task procedures is mediated by knowledge. Correlations between task knowledge and task performance were high, particularly for the two procedural tasks with the lowest skill requirements. The correlations reached as high as .75 in spite of the fact that the range of possible test performance seldom exceeded seven points. When performance was aggregated over tasks, the correlations tended to be more on the order of .80.

Substantial differences among methods of knowledge testing were not found. The conventional written multiple-choice test did essentially as well as the pictorially based methods in distinguishing masters from nonmasters. (In this connection, however, it should be noted that test questions were carefully directed at steps necessary in task performance, and did not include those marginally relevant knowledge items often found on such tests.) Failure of the Picture Sort tests to correlate higher with performance was an unexpected result. This method was designed to tap more fully all knowledge aspects of task performance, including recognition of the steps, their correct outcome, and sequence. In so doing, however, it may well have become the most demanding test technique from the standpoint of method-specific mediation requirements; that is, the examinee must first analyze what he does in performing the task, and then synthesize it a step at a time by sorting through a large number of pictures more or less representative of his mental images of the task. That kind of abstract manipulation probably taxes the intellectual and visualization abilities more than we originally anticipated. In support of this speculation, there was some indication that *S*s in the low mental ability group had more trouble with this test method than with others (Figure 1). The written and pictorial multiple-choice tests, though more dependent on literacy, represent a culturally familiar method. The Picture Outcome method appears to be the simplest in the sense of minimizing both literacy and method-specific mediational demands, and is certainly worthy of further study and development as an efficient method of knowledge testing.

Correlations between knowledge and performance were not significantly different for high versus low mental ability *S*s. Yet there was a slight but noticeable trend toward larger correlations within the low mental ability group. The possibility that knowledge measures—including the standard multiple-choice test—are better predictors of task mastery for those of below average mental ability is intriguing. If true, we need to reevaluate the popular notion that knowledge tests of manual performance are unfair to those less apt in the academic skills of reading, writing and symbol manipulation. The notion is probably valid, but it may be so for reasons quite different than normally offered. Knowledge tests apparently are good predictors of performance on low-skill procedural tasks among people of low to moderate mental ability. The unfairness lies not in the inability of this group to use a knowledge testing medium, but in the tendency of brighter people to over use it. The hypothesis here is that some minimum level of ability, whether innate or acquired, is necessary to handle the symbolic and semantic

demands of a knowledge test; but beyond that level, correlated factors such as test-wiseness begin to moderate the true relationship between task knowledge and performance. Two additional features of the data tend to support this speculation: a) higher average knowledge test scores for the high mental ability group, and b) relatively more false-positive errors in predicting mastery among this group.

If one were urged to recommend, on the basis of this study, a method of testing knowledge on low-skill procedural tasks, the Picture Choice would probably have to be named. The data are certainly not conclusive, but this method came the closest to meeting the overall validity criteria: it demonstrated a high correlation with hands-on task performance; the correlation was relatively constant over the range of mental ability; and, the distributions of classification error were more nearly proportional for the two levels of mental ability. Moreover, the Picture Choice method was second only to the hands-on test in examinee preference.

AN OVERVIEW OF THE SKILLS QUALIFICATION TEST DEVELOPMENT WORKSHOP

Charlotte H. Campbell, J. Patrick Ford
and Roy C. Campbell

HUMAN RESOURCES RESEARCH ORGANIZATION
Fort Knox, Kentucky 40121

Paper for:

Military Testing Association Conference
San Antonio, Texas

October 1977

OVERVIEW OF SQT DEVELOPMENT WORKSHOP

The Training Developments Institute (TDI), the Individual Training and Evaluation Directorate (ITED), and the Human Resources Research Organization (HumRRO), are conducting workshops that present the basic principles of developing criterion-referenced tests as the principles apply to developing a Skill Qualification Test (SQT). The workshops were developed by HumRRO under contract with the Army Research Institute (ARI).

In this paper I will describe the need for such a workshop, the constraints on the workshop, and the characteristics of the workshop.

NEED

The SQT is a highly complex system in form, development, and administration.

SQT is a criterion-referenced test; that is, performance on the test is measured against a standard determined in advance by an analysis of job performance requirements. Criterion-referenced testing, in various forms and under various names, has been around for a long time. But the SQT is, in form, development, and administration, different from other forms of criterion-referenced systems. The need for training, as provided by the workshop, arises from these differences and the problems associated with them.

In form, the principal difference is in the three components, or types of test mode, within an SQT. An SQT consists of:

- Hands-On Component (HOC)
- Performance Certification Component (PCC)
- Written Component (WC)

The Hands-On Component represents the most common type of criterion-referenced testing. The examinee performs a task, or part of a task, under standardized conditions, and is evaluated against a standard of job performance. Even though hands-on testing is most clearly criterion-referenced, an SQT may have no hands-on component.

The Performance Certification Component also resembles traditional criterion-referenced testing. The conditions under which the test is performed are not always standardized, but the allowable range of conditions is specified. A soldier's performance is usually evaluated by a supervisor as he performs the task on the job. The PCC is clearly criterion-referenced: performance is measured against a standard. But some SQT will have no PCC.

The third component of the SQT is the Written Component. In appearance, it resembles any other written test: there are item stems and alternative responses, and the examinee chooses the right response or responses. But where most written tests sample a domain of knowledges, the Written Component is criterion-referenced: the examinee performs portions of tasks, or indicates how the task is performed. The scorable units, of 2-10 items each, cover discrete tasks, and the item alternatives represent the actual alternatives that the soldier encounters on the job. The examinee then receives a score, GO or NO GO, for each task. All SQT will have a written component.

Thus the form of an SQT is complex. Problems arise in selecting tasks to be tested, adapting task analysis data to make them suitable for constructing of criterion-referenced

tests, and determining which of the three components is best suited for testing each task. "How we used to do it" sometimes interferes with how it must be done.

In addition to form of SQT, problems also arise in development. Development of the SQT involves much more than the construction of the test for each task. The primary problem in development arises from the requirement that an SQT must be validated. Unlike most criterion-referenced testing systems, an SQT must be tried out for reliability and validity before it can be administered in the field. For the HOC, the procedure involves checks of scorer reliability and test feasibility; for the PCC, scorer reliability, test feasibility, and systematic monitoring of testing must be ensured; for the WC, predictive and content validity are of concern. The procedures for validating an SQT are unique.

In the area of administration, the SQT also produces some distinctive requirements and problems. Everyone in a particular MOS skill level, worldwide, will take the SQT during the same test period. Besides being large scale, testing is also decentralized. Test Control Officers at each installation will conduct SQTs. For the HOC and PCC, this means that developers must prepare every precise performance measures, test conditions, and instructions to the Test Control Officers, scorers, and examinees. For the WC, the large scale, decentralized testing means that every response must wind up as a mark on a machine-scored answer sheet.

These unique characteristics of an SQT, in form, development, and administration, have created considerable problems. Even experienced test developers, even experienced criterion-referenced test developers, have found that developing an SQT is not an easy job. And at most Test Development Agencies (TDA), the SQT developers are not test experts, but subject matter experts. Their expertise is vital in SQT development, but not sufficient.

Early in the history of SQT development, ITED became aware of recurring problems on nine major tasks or aspects of the test developer's job:

- Select Tasks for Testing
- Review Task Analysis
- Allocate Tasks to Components
- Construct Hands-On Component
- Tryout Hands-On Component
- Construct Written Component
- Validate Written Component
- Construct Performance Certification Component
- Prepare SQT Notice

Guidance on the nine tasks was published as the *Guidelines for Development of Skill Qualification Tests.* In addition to this document, however, a need was perceived for a controlled, systematic approach to training and assisting individual developers in the implementation of the principles contained in the Guidelines. The SQT Development Workshop was proposed as a means to provide monitored practice in the skills involved in the nine tasks.

The overall objective for the Workshop is to prepare people at Test Development Agencies to perform these nine tasks, and to apply them to their own SQT development.

CONSTRAINTS

The workshop had to accommodate three constraints. The first of these is that it had to be exportable. While responsibility for development of the course was assumed by the U.S. Army Training and Doctrine Command (TRADOC), ultimately the

implementation of the course is the responsibility of the individual TDA. TDA traditionally experience considerable turnover among SQT personnel. TDA must be able to repeat the course as often as their needs dictate.

The decision was made to make the course a part of the total Faculty Development Program under the direction of TDI. At the TDA or school level, the course would be the responsibility of the Staff and Faculty Development section. The requirement then, was that the course be exportable to the extent that it could be taught to staff and faculty personnel who would then act as course managers at their TDA and conduct the course as needed to meet their own needs.

The second requirement was that the course be self-paced. While the TRADOC training philosophy incorporates self-pacing in its instructional model, this was not the only basis for this requirement. Persons who are assigned to SQT development have a variety of experience. Some know a great deal about testing but little about the practical limitations of SQT. When participants come to the workshop, their actual experience with SQT ranges from absolutely no prior exposure to SQT to two years working in SQT Development. Thus, as the need to learn about various aspects of SQT varies, the workshop had to allow individuals to work at their own pace.

The third constraint related to the time of the workshop. You may have heard that training should be limited only by the amount of time required for students to master the objectives. You may even have said it. But there is almost always an outside limit. For this workshop the limit is two weeks. Managers are just not willing to allow people to be away from their desks for more than ten days to learn to develop an SQT.

These constraints have been faced by other developers. As part of the total Faculty Development Program, TDI has successfully implemented a Criterion-Referenced Instruction (CRI) Workshop, developed by Mager Associates. This CRI workshop has become the basic foundation for a family of staff and faculty development programs which will provide the necessary in-house training capability in each TRADOC training facility.

According to the CRI model, the overall objective for a training program is broken down into subordinate objectives, and training is presented in modules corresponding to these subordinate objectives. Within some limits, participants choose the sequence in which they will tackle the modules. At the beginning of each module, the objective for that module is stated, the criterion test is described, and resource references for the material are listed. Each participant decides individually how much he must study and practice to pass the criterion test. A course manager monitors student progress, evaluates criterion tests, and serves as a learning resource when required by the student.

This basic framework was followed for development of the SQT Workshop.

CHARACTERISTICS OF THE WORKSHOP

The workshop is designed for worker-level development personnel, that is, the individual who actually must produce an SQT. Because of the detail in which the material is presented, it is not intended for senior or most middle management level personnel.

The workshop objective, to prepare people to apply the principles in the nine tasks, was broken out into 34 subordinate objectives. For example, one such subordinate objective (within the task, "Construct Hands-On Component") is: "Using a task analysis for a task allocated to the HOC, construct performance measures for process scoring, product scoring, or combination scoring." Thirty-four modules were prepared for these subordinate objectives. Each module contains explanatory text, examples and practical exercises. The examples and practical exercises are for the most part based on common military tasks. Sample tasks were chosen to illustrate the principles being discussed, and are intended to be familiar to most course participants.

For each module, there is also a criterion test. Each criterion test involves one of two types of material. In some tests, the material is standard, that is, the participant is given a situation, task, or other information, and applies the concepts put forth in the module to satisfy the requirements of the test. In others, the participant is expected to work with a task and material of his choosing from the MOS and skill level that he will be working with during SQT development.

The balance between the two types of material was not easy to achieve. Standard tests are amenable to very specific feedback, and make the role of the course manager easier. However, requiring the developer to use his own material helps to overcome the "My MOS is different" syndrome by showing developers the adaptability of the course materials. In this way, the participant develops a greater appreciation of the flexibility and relevance of the principles to his own job. Approximately one-half of the criterion tests involve developers in using their own tasks.

In the workshop, the nine major tasks discussed earlier were grouped into seven phases of skill development. These seven phases are necessary for complete development. Although emphasis in the workshop is on the individual modules, not the phases, in the time remaining I will briefly outline what is involved in each of the seven phases. (See Figure 1.)

The first phase, for all participants, is the analysis and planning phase. At the beginning of the workshop, participants select an MOS with which to work, one with which they are familiar. They begin with ten tasks from one skill level of that MOS. In one module, participants identify sources of information on each task that are objective indicators of need for evaluation. In another module, participants group the ten tasks according to the extent of known performance deficiencies. These modules lead participants to select for testing those tasks which promise the greatest payoff in testing. From the ten tasks, a course manager then selects five tasks with which the participant continues to work. Then, in the criterion test for the task analysis module, participants review and, if necessary, revise existing task analysis data for those five tasks to make them suitable for test construction. The final module in the analysis and planning phase covers allocating tasks to components. In the criterion test, participants assign each of their five tasks to the HOC, the PCC, or the WC. High skill physical tasks are allocated to the HOC or the PCC; mental tasks and low skill physical tasks are allocated to the WC.

After participants finish the analysis and planning modules, they branch into either the HOC construction phase or the WC construction phase. During the construction phases, participants work with the tasks selected earlier. For the HOC construction, there are modules for some preliminary decisions called for in the Guidelines. Then they work on modules which require that they construct two complete hands-on scorable units, to include performance measures, conditions, examinee instructions and scorer instructions.

The WC construction phase also requires participants to write scorable units for tasks they selected. They practice constructing two kinds of written test: written performance tests, which require examinees to perform part or all of a task, and performance-based tests, which require examinees to answer questions about how a task is performed.

After participants finish the construction phase for a component, they move to the validation phase for that component. Here, the activities and criterion tests are standardized, and address the analysis of data and revision of scorable units based on validation results.

The HOC validation procedure checks interrater reliability, acceptability, and feasibility. The modules cover locating faults based on a tryout with experts, computing scorer agreement, checking feasibility of a scorable unit, constructing a station-load table, and revising hands-on scorable units.

The WC validation procedure checks discriminant validity and acceptability. Three options for validation are available, based primarily on the number and types of soldiers

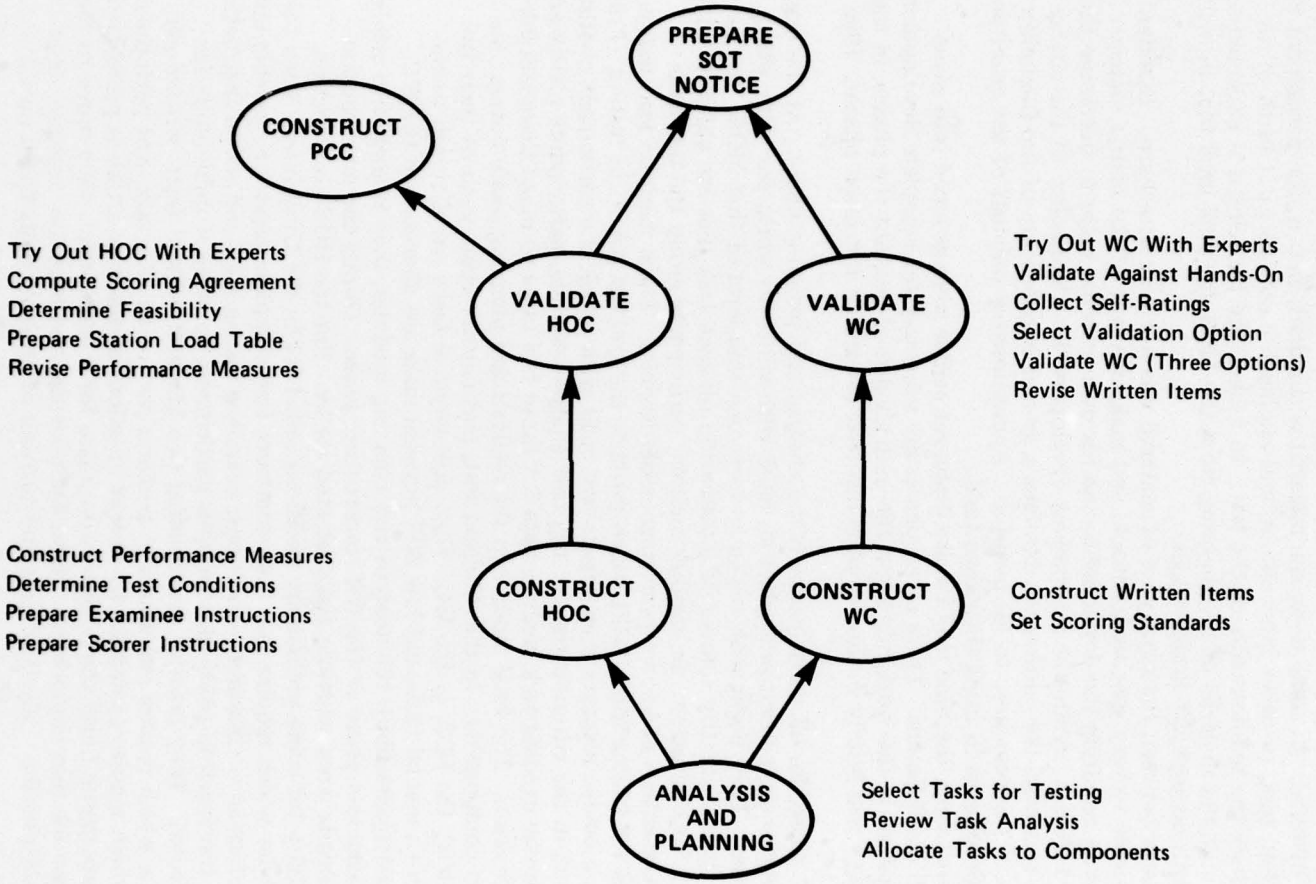

Figure 1. **Phases of SQT Development Workshop**

to which the developer has access. The validation modules cover collecting self-ratings, locating faults based on a tryout with experts, validating written scorable units against hands-on tests, selecting a validation option and analyzing data on each of the three options, and revising written scorable units.

The revision modules cap each validation phase. ITED's policy in regard to validation is that the results give a basis for locating and correcting faults in a test. The modules present troubleshooting charts for hands-on or written tests. The activities and the criterion tests present summaries of results that indicate malfunctions on given items or performance measures. Using the troubleshooting charts, participants then modify the test to correct the probable causes of the malfunction. These modules call for additional practice of the skills acquired during the construction phases.

The sixth phase, dealing with the PCC, focuses not only on the procedures for constructing the PCC but also on procedures for validating and monitoring it. Participants again work with one of their own tasks. They describe how the test will be conducted, how it will be validated, what kinds of results would indicate units for followup checks, and how the checks will be conducted.

In the final phase, after participants have developed a scorable unit for each component, they prepare an SQT Notice. This is primarily a check on their mastery of the format for the Notice. It also provides a neat wrap-up of the course.

In this way, workshop participants work through a full cycle of SQT development in about ten days. Workshop materials will be revised based on our experience with 15 TDA currently being trained. Reaction from those who have already received the training has been overwhelmingly favorable. Even individuals who have had no previous contact with test development or with SQT have expressed confidence in their ability to fit into the SQT system after taking the workshop. Likewise, participants, who have had prior work with SQT state that the workshop has improved their skills and capabilities. The revised workshop will then be added to the Staff and Faculty Development courses at the TDA. The concentrated practical work in this workshop will thus be part of the on-going TRADOC support of TDA to accomplish the unique goals of SQT.

A PAIRED-COMPARISON APPROACH FOR ESTIMATING TASK CRITICALITY

James H. Harris, William C. Osborn
and John A. Boldovici

HUMAN RESOURCES RESEARCH ORGANIZATION
Fort Knox, Kentucky 40121

Paper for:

Military Testing Association Conference
San Antonio, Texas

October 1977

A PAIRED-COMPARISON APPROACH FOR ESTIMATING TASK CRITICALITY

Training resource limitations demand that choices be made about what to include in training, and what to exclude. Agreement seems widespread that training programs should minimally include tasks that are critical to effective job performance and cannot be performed by new trainees. In military training contexts, this reduces to including in training those tasks that are critical to effective performance in combat. Since combat cannot be realistically simulated, a measurement problem immediately arises; namely, how to measure task criticality.

Prescriptive training development literature typically mentions task criticality as an important consideration in determining training content. The literature is, however, vague on the question of how to measure criticality, and silent on the measurement issues associated with criticality estimation.

Conventional training development methods deal with the problem of selecting tasks for inclusion in training in the following way: A job analysis is conducted, resulting in a task list or "inventory." Expert judgment is then used to rate the criticality of each task, usually on some *n*-point scale ranging from "irrelevant to the job" to "highly critical to mission accomplishment." The tasks receiving the highest ratings are selected for inclusion in training, and those receiving low criticality ratings are excluded or deemphasized. Since the content of training frequently is determined on the basis of criticality ratings, a question arises as to how much confidence can be placed in the ratings. One index of that confidence is inter-rater reliability: to the extent that several raters independently produce similar criticality ratings, confidence in the job-relevance of training content based on the ratings increases. The test-development axiom is directly analogous: reliability is necessary for validity. Applied to training content, the axiom becomes "reliability (of criticality ratings) is necessary for job-relevance (of training content)."

The reliability of criticality ratings that are used for determining training content seldom is reported (McCluskey, *et al.*, 1975; McKnight and Hundt, 1972). In the few instances where reliability has been reported (Ammerman and Pratzner, 1975) rater agreement has been poor—too low in fact for the ratings to be of practical use. We suspected that low-reliability in these studies was due to two important factors. First, tasks were rated on an absolute rather than comparative basis which, among other things, tends to restrict the range of ratings. Second, there were no definitions of criticality to restrict the dimensions along which raters' judgments were to be made. Therefore, we decided to try a rating technique that would:

1. Enable raters to compare tasks to one another rather than to a numerical scale.
2. Simplify the judgment process.
3. Provide an operational definition of criticality.

The paired-comparison technique, although used to scale a variety of kinds of stimuli from things to people, has not been used, to our knowledge, in rating job tasks. In the paired-comparison approach, raters are presented two stimuli and asked to judge which stimulus is greater with respect to some characteristic such as size, brightness, beauty.

We have tried the paired-comparison technique in two studies (Boldovici, *et al.*, 1976; Boldovici, *et al.*, 1977). In the first project, two-hundred forty tank gunnery tasks were ranked in terms of criticality, which was determined by the use of the paired-comparison technique. The Tank Commanders serving as respondents were presented with many pairs of target/range combinations. (An example of a pair of target/range combinations is tank at 2000 to 2500 meters, and light-armored vehicle at 500 to 1000 meters.) The respondents were instructed to assume that they had encountered each pair of target/range combination on the battlefield, and that they could not engage the targets simultaneously. They were then asked to indicate which one of the two target/range combinations that comprised each item they would engage first. A criticality score was computed by counting the number of times each combination was chosen as more threatening ("would be engaged first") and dividing by the number of times it could have been chosen (Guilford, 1954). Inter-rater reliability was in the high nineties. Since the rated items varied only in target type and range, the judgments about target threat or criticality were easy to make. The high degree of rater agreement probably also reflected certain learning experiences that the subjects had in common: Tank Commanders receive formal training in assessing target threat. The high inter-rater reliability, therefore, may simply have indicated that all of the subjects had learned "the same things." The second project provided for answering the question whether similarly high inter-rater reliability could be achieved using the paired-comparison technique with a less homogenous sample of armor tasks, where the dimensions for making the criticality judgments were less obvious than target or range, and where the respondents had not received formal instruction in making judgments of the kind required for the ratings.

Forty-eight captains, who were enrolled in the Armor Officers' Advanced Course (AOAC) at Fort Knox during the conduct of the project, served as respondents. Twelve forms of a paired-comparison questionnaire were used. The stimuli to be rated in each form were the tasks for one of four crew positions (Driver, Loader, Gunner, Tank Commander) in one of three tanks (M60A1, M48A5, M60A3). The design of each form of the questionnaire can be illustrated by describing how the form for the M60A1 Driver tasks was designed. Seventy M60A1 Driver tasks were identified during the task-description part of the project. The number of possible different pairs of 70 tasks, then, is 70 x 69/2 = 2415. This, of course, would have been too many judgments for each respondent to make. A partial paired-comparison design (McCormick and Bachus, 1952) was used, in which each of the 70 tasks was paired with each of seven other tasks. This partial pairing approach yielded 245 unique pairs of tasks for the M60A1 Driver. The numbers of pairs of tasks for the other 11 forms of the questionnaire ranged from 135 to 280.

The respondents were instructed to assume that they were company commanders choosing crew members to take on a mission in which fire would be exchanged with the enemy. They were then asked to indicate which of two crew members they would choose, based on whether the crew member could do one or the other of a pair of tasks. An example of a pair of tasks for the M60A1 Driver is:

1. Start tank engine.
2. Move vehicle into defilade firing position upon enemy contact.

Criticality values were calculated for each of the twelve sets of tasks by a standard three-step procedure (Guilford, 1954) which placed the twelve sets of values on a similar positive scale. Inter-rater reliability was estimated by correlating scale values for tasks common to the three tanks. The correlations ranged from .55 to .79, with an average of .68. All were statistically significant ($p < .05$).

The paired-comparison technique holds promise as an approach for estimating the relative criticality of tasks. However, the inter-rater reliability estimates and questions

about the validity of the results obtained in the two projects raise separate issues for discussion regarding how to generate task criticality estimates that are reliable and valid.

Reliability

The reliability of the criticality estimates obtained in the second paired-comparison study, though statistically significant and probably greater than the reliabilities of criticality ratings in studies using absolute ratings (Harris, *et al.*, 1975), seems only marginally acceptable, particularly when compared to the results of the first paired-comparison project. The earlier project, however, differed from the later one in several respects which give rise to some tentative operating assumptions on how to generate criticality estimates that are highly reliable. The reliability of the criticality ratings can be expected to increase with:

1. Specificity of the dimensions along which criticality ratings are to be made. To the extent that investigators can create a uniform set among raters as to the dimensions along which judgments are to be made, rater agreement should increase. Without clear specification of the dimensions for making judgments, raters will "make up" their own dimensions. And if these dimensions differ from one rater to the next, rater agreement will suffer.

2. Common learning experiences among raters. The obvious recommendation—that raters should practice making judgments of the kind required by the criticality study—is warranted only when the condition just discussed (specific dimensions) is met. Practice might otherwise simply reinforce idiosyncratic rater behavior and thus reduce rater agreement.

3. The extent to which complete pairings of the tasks to be rated is approximated. The desirability of eliminating the "luck of the draw" in determining which tasks get paired with one another must, however, be traded off against the heavy rater workloads that characterize complete pairings with large numbers of stimulus materials.

4. The number of times each stimulus is rated. Every respondent need not rate every possible pair of tasks, though this may be desirable. Decreasing the workload of each subject can be accomplished in several ways. Partial pairings can be used, with all subjects rating all pairs. Or complete pairings can be used with some of the subjects rating some pairs and not others. Various mixes of the approaches also may be used—partial pairings with some subjects rating some pairs and not others. The optimal compromises are unfortunately, not known. Examinations would be interesting, of the effects on rater agreement of various reductions (combined and in isolation) in number or proportion of compared pairs, number or proportion of raters rating each pair, and number of observations per stimulus and pair.

> The generality of the results of such research would, of course, never be fully established. Questions would always remain about the effects of stimulus materials, instructions to raters, rater experience and so forth, on the results obtained. But if confidence is desired in the results of studies that purport to measure the criticality of combat tasks, then additional research on factors affecting rater reliability seems necessary.

Validity

Any study which claims to measure task criticality raises questions associated with the construct, content, and predictive validity of the results obtained. Construct validity is concerned with the extent to which one measured what one intended to measure. Instructions to the respondents should be designed to create a set for judging criticality and criticality alone. But raters' judgments may be influenced by extraneous considerations such as how difficult a task is to learn or perform, or how frequently it is performed on the job. Questions about construct validity will remain as long as reasonable counter-interpretations of the results can be advanced (Cronbach, 1976).

Content validity addresses the extent to which items used in questionnaires represent the universe of items. The issue of how well the universe of subject matter is sampled can never be fully resolved. Resolution would require widespread agreement on the adequacy of the descriptors used to define the universe, and on precise definitions of what constitutes adequate sampling. On the other hand, if a job domain is carefully partitioned into tasks, and all tasks are included in the criticality study, content validity is not a major concern.

Predictive validity is concerned with to what extent would the criticality scores or predictions made from them, correlate with a direct measure of criticality. Establishing the predictive validity of the results of a criticality study would require correlating the obtained criticality scores with a direct measure of criticality. Obtaining direct measures of task criticality in combat is, of course, out of the question. Intermediate criteria, combat simulations, for example, might be used in studies of predictive validity. Of course, achieving adequate measurement reliability under simulated combat conditions would be very expensive, though absolutely essential if any important decisions are to be made based on the simulation results.

Concern with the validity of the ratings, though appropriate, may be premature. Reliability issues associated with estimating the criticality of job tasks have only begun to be raised. Given a) that nothing is known about the validity of criticality estimation, and b) choices between results of known and unknown reliability, training developers would seem well advised to use results whose reliability is known. In this respect, it appears that the paired-comparison technique holds promise as a method of rating task criticality.

REFERENCES

Ammerman, H.L. and Pratzner, F.C. *Occupational Survey on Auto Mechanics: Task Data from Workers and Supervisors Indicating Job Relevance and Training Criticalness.* Columbus, Ohio: Ohio State University, 1975.

Boldovici, J.A., Wheaton, G.R., and Boycan, G.G. *Selecting Items for a Tank Gunnery Test* (Draft). Fort Knox, Kentucky: Human Resources Research Organization (HumRRO), 1976.

Boldovici, J.A., Harris, J.H., Osborn, W.C., and Heinecke, C.L. *Criticality and Cluster Analyses of Tasks for the M48A5, M60A1, and M60A3 Tanks.* Fort Knox, Kentucky: Human Resources Research Organization (HumRRO), 1977.

Cronbach, L.J. Test Validation. In R.L. Thorndike, (Ed.) *Educational Measurement (Second Edition)*, Washington, D.C.: American Council on Education, 1976.

Guilford, J.P. *Psychometric Methods.* New York, New York: McGraw-Hill, 1954.

Harris, J.H., Campbell, R.C., Osborn, W.C., and Boldovici, J.A. *Development of a Model Job Performance Test for a Combat Occupational Specialty. Volume 1. Test Development.* Fort Knox, Kentucky: Human Resources Research Organization (HumRRO), 1976.

McCluskey, M.R., Jacobs, T.O., and Cleary, F.K. *Systems Engineering of Training for Eight Combat Arms MOSs*, Alexandria, Virginia: Human Resources Research Organization (HumRRO), 1975.

McCormick, E.J. and Bachus, J.A. Paired comparison ratings. I. The effect on ratings of reductions in the number of pairs. *Journal of Applied Psychology*, April 1952.

McKnight, J.A. and Hundt, A.G. Driver Education Task Analyses: *The Development of Instructional Objectives.* Alexandria, Virginia: Human Resources Research Organization (HumRRO), 1972.

RELIABILITY IN MEASURING UNIT PERFORMANCE

John A. Boldovici, William C. Osborn and
James H. Harris
HUMAN RESOURCES RESEARCH ORGANIZATION
Fort Knox, Kentucky

Paper for:
Military Testing Association Conference
San Antonio, Texas

October 1977

RELIABILITY IN MEASURING UNIT PERFORMANCE

A central problem in all evaluations, and especially in evaluations of combat units, is how to incorporate characteristics of good measurement in the evaluations. Characteristics of good measurement include comprehensiveness, cost-effectiveness, validity, and reliability. Our concern in this paper is with reliability; for without reliability, comprehensiveness is of little value, and cost-effectiveness and validity cannot be achieved.

Reliability refers to the extent to which:

1. Two or more independent observers produce similar results, and
2. Measures of an event taken at one time are identical to measures of the same event taken at another time.

The performance of combat units, at least in the Army, is increasingly being evaluated in the context of large-scale, free play simulated combat exercises. The ARTEP (Army Training and Evaluation Program) is an example. The results of performance evaluations in simulated combat are used by policy makers in decisions about training needs and combat readiness. Given the importance of decisions about training needs and combat readiness, and given the dependence of these decisions on unit performance evaluations, a question naturally arises as to how to maximize reliability in measuring unit performance.

PURPOSE

The purpose of this paper is to present hypotheses about variables that affect the reliability of unit performance measurement, and to outline research for testing the hypotheses.

SOURCES OF MEASUREMENT RELIABILITY

Measurement can be viewed as consisting of three phases:

1. Observer Preparation
2. Observation
3. Recording and Reporting

Variables that affect measurement reliability are at work within each of the three phases of measurement—variables that affect the extent to which two or more observers produce similar measurement results, and the extent to which measures taken at a given time are representative of measures taken at another. Hypotheses about the variables in each of the three measurement phases follow.

Observer Preparation

Reliability of measurement will increase with the consistency or uniformity of understanding among observers about the rules of observation and recording. Ideally,

observers should be standardized, and measures should be taken to assess the degree to which they have been standardized. Measurement reliability may be increased by manipulating the following variables in the observer preparation phase:

1. Specificity of instructions. Reliability is likely to be greater when the instructions to observers are highly specific than when instructions are general and loosely stated.
2. Timing of instructions. Instructions to observers should not be given so far in advance of observation as to permit forgetting, or so late as to preclude learning.
3. Practice in observing and recording. Measurement reliability will be greater when observers have practice measuring and recording the events of interest than when they have not. The practice variable interacts with timing of instructions, in that instructions to observers should be given far enough in advance of observation to allow time for practice.
4. Testing observers. Measurement reliability can be indirectly increased by the use of tests to make sure that observers are capable of performing whatever measurement operations will be required of them.

Observation

Even with very careful observer preparation and totally standardized observers, measurement reliability will be affected by variables at work during the observation (measurement) process.

Properties of the events or things to be measured can affect measurement reliability. Measurement of unidimensional events will, for example, be more reliable than measurement of multidimensional events (all other things being equal). This is related to perceptual "clutter," or limits on observers' information-processing abilities. Within rather broad limits, observers who are asked to make large numbers of simultaneous observations and measures will produce less reliable results than will observers making smaller numbers of observations.

Another property of the events or things to be measured that affects measurement reliability is stability (or its opposite, transience). The results of measuring the diameter of a wooden ball will, for example, be more reliable than will the results of measuring a mercury "ball"—once again, all other things being equal.

Other properties of events to be measured that will influence reliability are time-sharing, noise, and "observability"; that is, measurement reliability may be expected to decrease with the extent to which the observed event is:

1. Time-shared with other events.
2. Embedded in noise.
3. Not directly observable.

Strategies, rules, and procedures for measurement also affect reliability. Observers may be expected to perform more reliably, for example, to the extent that they are:

1. Required to make comparative rather than absolute judgments.
2. Given a well defined standard stimulus.
3. Alerted as to what to observe (anticipate likely errors).

4. Given the opportunity to observe an event more than once.
5. Given scoring aids or templates.
6. Required to measure only, and not process measurement results.

Recording and Reporting

Even with adequate observer preparation and careful control of the measurement process, measurement reliability will be affected by variables operating during the recording and reporting of measurement results. These variables include:

1. Timing. Measurement reliability will increase with decreased time between observation of the event of interest and recording of results.
2. Design of recording forms. Well designed data recording forms minimize the amount of judgment and decision-making required for their use, and thereby increase the reliability of recorded results. Simplicity in data-recording forms, for example, minimize data-recording time, and therefore allows more time for observation.

Unit performance measurement probably is unreliable because of the influence of all of the variables mentioned above. These variables serve to decrease the reliability of operations as simple and straight-forward as measuring length with a ruler. The considerable complexity of free play simulated combat guarantees that measurement reliability problems will be great.

In the observer preparation phase, for example, observers may not be standardized for any number of reasons. Instructions for measurement may be too general, and may not be given at the right time. Observers may not have enough practice to permit performing their measurement duties in accordance with the intent of the test designers. And practical constraints (e.g., time, money) may preclude ascertaining whether observers are capable of performing their measurement duties before "turning them loose."

In the observation phase, observers may be required to make simultaneous judgments along more dimensions than their sensory apparatus can comfortably handle. The measurement instruments may permit too much subjectivity and expertising. Strategies for measurement may be inappropriate (single rather than multiple observations, for example). And the nature of the required judgments and decisions may invite unreliability.

In the recording and reporting phase, unreliability may be promoted by the length of time between observation and recording of results, and by formats for recording results.

The possible influences of the variables discussed above demand that research be undertaken on methods for improving the reliability of unit performance measurement, for measurement without reliability will lead to wrong decisions about training needs and readiness.

PHOTOGRAPHY AND MEASUREMENT RELIABILITY

The conduct of measurement reliability studies requires that whatever is to be observed and measured (simulated combat, for example) must:

1. "Sit still" long enough to permit observers to make the required measures.
2. Be presented uniformly or varied systematically for various groups of observers.

These two requirements, and the high cost of field studies using simulated combat, make the conduct of field studies of measurement reliability impractical. The requirements for "sitting still," for uniform or systematically varied presentation, and for low cost can be met by the use of photography.

Motion pictures of simulated combat can be made, using real combat vehicles or models. Models seem preferable for two reasons. The first is low cost. The second is that research on reliability of measuring unit performance does not require perfect fidelity or realism in the events to be observed and measured. As noted earlier, the main requirement is for a set of events that can be presented uniformly to various observers, or varied in accordance with requirements of the experimental design.

Subtle errors in tactics and operations can be deliberately incorporated into motion pictures, for the purpose of producing variability in observers' response to events presented in the film. And by editing videotape versions of the film, the amount of information available to various groups of observers can be systematically varied.

Studies of reliability in unit performance measurement should take the following general form: A set of events is selected for observation and measurement (e.g., a part of the ARTEP). Several groups of subjects view the events, observing, measuring, and evaluating according to instructions and experimental conditions. Systematic variations are introduced in variables in any or all of the three phases of measurement. As implied earlier, variations could be introduced in the kinds of instructions given to observers, the specificity of the instructions, amount of practice given to observers, kinds of instruments and measurement strategies, and so forth. In all cases the dependent variable is an index of inter-observer reliability; e.g., a simple "percent-agreement" score to indicate the extent to which observers produce similar results measuring the same things. Variables that affect reliability are identified, and can be incorporated into "how-to" literature for reliable unit performance measurement.

The conduct of research along the lines suggested above seems warranted, because the results would lead immediately to action recommendations for improving measurement reliability, and could be incorporated directly into any program for measuring unit performance.