ARI Research Note 95-12

Optimal Averaging in Performance Tests

Marshall B. Jones
Pennsylvania State University

Research and Advanced Concepts Office
Michael Drillings, Acting Chief

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

Edgar M. Johnson
Director

Research accomplished under contract
for the Department of the Army

Pennsylvania State University

Technical review by

George Lawton

NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE 1995, January	3. REPORT TYPE AND DATES COVERED FINAL 5/86 - 9/89

4. TITLE AND SUBTITLE
Optimal Averaging in Performance Tests

5. FUNDING NUMBERS
MDA903-86-C-0145
0601101A
B74F
2901
C16

6. AUTHOR(S)
Marshall B. Jones (Pennsylvania State University)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Pennsylvania State University
Milton S. Hershey Medical Center
Hershey, PA 17033

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-BR
5001 Eisenhower Ave.
Alexandria, VA 22333-5600

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARI Research Note 95-12

11. SUPPLEMENTARY NOTES
COR: Michael Drillings

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words):

The purpose of this research was to develop a methodology for optimizing the temporal stability and predictive validity of performance tests and to apply that methodology to the Project-A, computer-administered tests. In the present research, a performance test is treated as a task to be practiced, and tests are analyzed as individual differences in skill acquisitions and retention. Classical test theory is also used. The predictive validity of the Project-A, computer-administered tests for a simulated anti-aircraft criterion task was studied over a 4-month interval in a sample of 102 college students; the 4-month temporal stability of the tests was studied concurrently in the same sample. Three of the 10 Project-A tests (Choice Reaction, Target Tracking 2, and Cannon Shoot) show a forward stability optimum. Cannon Shoot also has high predictive validity (.59). It could have the highest predictive validity of any test in the Project-A battery if its temporal stability could be improved. In none of these tests, however, can temporal stability be improved by lengthening the tests.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Performance test Optimal averaging	32
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std 239-18 298-102

OPTIMAL AVERAGING IN PERFORMANCE TESTS

OBJECTIVES

The purpose of this research was to develop a methodology for optimizing the temporal stability and predictive validity of performance tests and to apply that methodology to the Project-A, computer-administered tests.

APPROACH

Performance tests differ fundamentally from knowledge tests. In a knowledge test the order in which items are administered has minor effects and is usually ignored. In a performance test it is difficult or impossible to prevent knowledge of results and, as a result, order of administration matters greatly. In the present research a performance test is treated as a task to be practiced and test results are analyzed as individual differences in skill acquisition and retention. Classical test theory is also used.

PROCEDURES

The first step in the analysis is "forward averaging." It begins with the subjects' scores on the first trial. Then each subject's scores on the first two trials are averaged, then the first three trials, then the first four, and so on until the last trial is reached. This series of averages is then correlated with the corresponding series of averages from retest or a criterion to be predicted. Oftentimes temporal stability or predictive validity increases up to an optimal average and then decreases. When such an optimum is encountered, it means that the test's temporal stability or predictive validity cannot be improved by lengthening the test. It also means that the test can be shortened (back to the optimum) without loss of stability or validity. If a forward optimum is not encountered, forward averaging provides a basis for estimating what the effects of lengthening the test on stability or validity would be.

Backward averaging is the reverse of forward averaging. It begins with the last trial of practice. Then each subject's scores on the last two trials are averaged, then the last three trials, and so on until the first trial is reached. When backward and forward optima both occur, a simple algorithm is specified for determining the single most predictive average of consecutive trials.

RESULTS

The predictive validity of the Project-A computer-administered tests for a simulated Anti-Aircraft criterion task was studied over a four-month interval in a sample of 102 college students; the four-month temporal stability of the tests was studied concurrently in the same sample. Three of the ten Project-A tests (Choice Reaction, Target Tracking 2, and Cannon Shoot) show a forward stability optimum. Cannon Shoot also has high predictive validity (.59). It could have the highest predictive validity of any test in the Project-A battery if its temporal stability could somehow be improved. In none of these three tests, however, can temporal stability be improved by lengthening the tests.

The predictive validity of all ten tests can be improved by optimal averaging, in six cases significantly so. Overall, optimal averaging improves the predictive validity of tests with validities representative of real-world, job-performance validities from .34 to .45. Gains of this magnitude would be of major practical importance, especially since they can be obtained at no cost in test modification or testing time. These results, however, need to be cross-validated.

OPTIMAL AVERAGING IN PERFORMANCE TESTS

CONTENTS

LIST OF TABLES

LIST OF FIGURES

OPTIMAL AVERAGING IN PERFORMANCE-TESTS

INTRODUCTION

The theoretical problem of performance testing

The distinction between knowledge and performance testing turns on what one is trying to measure. A knowledge test samples what a subject knows, a performance test what he or she can do. Plainly, this distinction is not absolute. A mathematics test, for example, may involve not only what a subject knows but also what he or she can do with that knowledge. A memory search task may be facilitated if a subject has seen an unusual symbol before and knows what it is, say, a Greek omega. Nevertheless, most tests fall lopsidedly into one category or the other.

In a knowledge test the subject does not usually know whether he or she is right or wrong. As a result practice effects are limited to auxiliary aspects of the test (test-taking skills) and, while they exist, are not large (Messick & Jungblut, 1981; Wing, 1980). In a performance test, however, it is usually not possible to prevent the subject from obtaining some idea as to how well or poorly he or she is doing. As a consequence, subjects tend to do better on a test the more times it is administered to them (Bittner et al, 1983; Kennedy et al, 1981). In effect, each test administration becomes a trial of practice.

Psychometric theory is based on knowledge tests. The unit of analysis is an item and the order of administering the items is arbitrary. In performance testing, however, the unit of analysis is a trial and order of administration is not only nonarbitrary but often the only thing that distinguishes one trial from another. In a knowledge test it is not unreasonable to suppose that mean performance and interitem correlations are independent of order of

administration. In a performance test it is. Typically, performance improves with practice and intertrial correlations fall into a definite pattern as a function of order: the superdiagonal form (Jones, 1962).

The consequences of these differences for theory are drastic. It has long been known, for example, that intertrial correlations, unlike interitem correlations, may yield spurious results when subjected to conventional factor analysis (Humphreys, 1960). Also, the familiar formulae for adjusting reliability and validity for test length assume that average interitem (intertrial) correlation, $\bar{r}$, does not change with test length. In a superdiagonal form, as will be seen below, $\bar{r}$ definitely does change with test length. As a result, the Spearman-Brown and related formulae (Gulliksen, 1950) have to be reworked and reinterpreted if their use in performance testing is to be helpful and not misinformative.

The practical problem of performance testing

During the Second World War performance testing based on electromechanical apparatus (rotary pursuit, complex coordination, two-hand tracking, and the like) was widely and successfully used in military selection, especially for pilot training (Melton, 1947). The equipment, however, was heavy, bulky, difficult to maintain, and more difficult to replace. By the late 1950s all three military services had abandoned performance testing in favor of paper-and-pencil tests exclusively. Then in the late 1970s the advent of microcomputers reopened the possibility of performance testing, this time with equipment that occupied little space, did not break down frequently, and was easily replaced when it did. At the same time experimental psychology was undergoing a revolution of its own, as the discipline's central focus shifted from learning theory to cognition and

information-processing. The joint effect of these two developments was a new generation of cognitively oriented, microcomputer-based performance tests. The computer-administered tests in Project A are cases in point.

Unfortunately, all has not been clear sailing for this new generation of performance tests. The most serious problem has been that many tests have low reliabilities (Kyllonen, 1985). Predictive validities against real-world criteria are still sparse, but it seems likely that oftentimes they will also be low. An appropriate response to these difficulties involves more than making and trying out new tests. What is needed is a theory of performance tests, that is, an approach to test construction and validation that recognizes and capitalizes upon the distinctive properties of performance tests.

APPROACH

Superdiagonal form is one of the best established regularities in human learning (Jones, 1962, 1969). It refers to the essentially universal tendency for trials of practice to correlate more strongly the closer they are together in the practice sequence. Table 1 presents a hypothetical example. The correlation between neighboring trials is .80. If the two trials are separated by one intervening trial, the correlation drops to .65. If two trials intervene, the correlation drops to .50. The weakest correlation is between the first and last trials in the sequence, in the example, .05. In a typical motor-skills experiment, where each data point represents as much as 20 mins of practice, the superdiagonal pattern is always present and, in large samples, usually quite regular. In correlations among individual trials of practice, as in performance testing, the pattern may be very irregular.

Almost always, however, if correlations are averaged over groups of consecutive trials, the pattern can still be seen.

Table 1 illustrates another point, this one directly relevant to performance testing. In conventional test theory the Spearman-Brown (S-B) formula (Gulliksen, 1950) states that the reliability of a test i units in length*

$$R_i = \frac{i R_1}{1 + (i-1) R_1},$$

Table 1. Hypothetical correlations among seven trials of practice, together with the average correlation ($\bar{r}_i$) and reliability (R_i) as calculated by the Spearman-Brown formula up to a given trial.

Trial	Trial 1	2	3	4	5	6	7
1	-	.80	.65	.50	.35	.20	.05
2		-	.80	.65	.50	.35	.20
3			-	.80	.65	.50	.35
4				-	.80	.65	.50
5					-	.80	.65
6						-	.80
7							-
$\bar{r}_i$	-	.80	.75	.70	.65	.60	.55
R_i	.800	.889	.900	.903	.902	.900	.895

*As given here, the S-B formula assumes that all trials have the same variance. This restriction can be relaxed by restating the formula in terms of variances and covariances. For clarity of presentation, however, I will continue to use the simpler and more familiar form.

where R_1 is the reliability of a test of unit length. When $i \geq 2$, R_1 is taken as the average correlation among the i units, that is, $\bar{r}_i$. The first row at the bottom of the table shows this average correlation for the first two trials, the first three, out to all seven trials. As is clear from the table, these averages decrease as one moves forward from the first to the last trial. Since the correlations decrease along any row to the right, each new trial adds to the average a column of correlations lower than those already in it; hence $\bar{r}_i$ drops a notch.

Low reliability in a knowledge test is corrigible. It may be laborious to do, but in principle one can always lengthen the test, while maintaining the same average inter-item correlation, and thereby improve its reliability. In a performance test, however, $\bar{r}_i$ does not remain the same as the test is lengthened; it decreases. The bottom row in Table 1 gives R_i as calculated by the S-B formula for $i = 1, \ldots, 7$. As i increases, $\bar{r}_i$ both decreases and is more strongly amplified by the S-B formula. The amplification, however, is negatively accelerated while, in this example, the decrease in $\bar{r}_i$ proceeds at a constant rate. The upshot is that R_i increases sharply at first, reaches a maximum (at $i = 4$), and then decreases gently. In this case, therefore, reliability would not be improved by lengthening the test. In fact, the test could be shortened to 4 trials with no loss of reliability.

The superdiagonal pattern in Table 1 is perfectly regular, that is, constant within any given diagonal and regularly decreasing between diagonals. As we have seen, however, it nevertheless tends to yield reliabilities that increase to an optimum and then decrease gently. This tendency may be reinforced by other considerations. As the number of trials increases, some subjects may become fatigued or lose concentration. In addition, performance

in the presence of fatigue and wavering attention tends to be fitful and erratic. These changes introduce novel variance not present in earlier trials of practice. The effect is to produce a drop in correlational level late in practice and, therefore, to bring about a forward optimum earlier than it would have occurred in a perfectly regular pattern.

In practice, reliability as calculated by the S-B formula from a series of acquisition (test) trials is less interesting than temporal stability--that is, the correlation between test and retest over appreciable periods of time (months or years), where a subject's score is his or her average performance over the first i trials at test or retest. In a single test series a perfectly regular superdiagonal pattern, like the one in Table 1, suffices to produce a forward optimum (that is, a maximum prior to the last trial), provided the gradient away from the superdiagonal is steep and the series is long enough. When the subjects are tested in two well-separated series of trials, the conditions are somewhat different. Specifically, in the square array of correlations between test and retest the gradient along the rows to the right must be steeper than that up the columns. In remains true, however, that in stability as well as in reliability certain variants of superdiagonal pattern are sufficient to produce a forward optimum. Furthermore, these variants may be brought about by effects such as fatigue or loss of concentration.

Forward averages may also be correlated with an external criterion. When they are, the correlations (predictive validities) always rise at first and sometimes reach an optimum, after which they decrease. It has long been recognized that abilities may change with practice (Fleishman & Hempel, 1954; Ackerman, 1987) and that, as they do, the relation of the practiced test to an

external criterion may also change. There is, therefore, no reason to be surprised if the average of the first $\underline{i}$ trials sometimes predicts a criterion better than the average of all trials given.

Forward optima in temporal stability and predictive validity have important implications for the construction and validation of performance tests. If a test shows a forward optimum in stability, the implication is that lengthening the test will not improve its stability. It is true that if the test were lengthened, stability, after decreasing for a stretch of trials, might start increasing again to a second optimum. To date, however, I have not seen any such second increase. One does see small departures from an increasing, level, or decreasing curve but not a second increasing trend in the curve's general direction. If, however, an optimum once reached will not be exceeded or, in the worst case, not exceeded by much, then lengthening the test will not improve its temporal stability.

If a test has not reached an optimum or asymptote with the number of trials given, it is possible to project where the optimum would fall if the test series were lengthened. This projection is based primarily on extrapolating the course of $\bar{r}_{ii}$, the average correlation between test and retest trials up to test or retest trial $\underline{i}$, from the existing series to next and following trials. Such a projection is, of course, no better than the extrapolation on which it is based. Still, forward averages provide an empirical basis for decisions regarding the length of a performance test. It can tell us when a series of trials is already long enough, whether it might be shortened without loss of stability, how much it would have to be lengthened to reach an optimum, and how much of a gain could be realized by so lengthening it.

Once a test has been constructed, it may be used to predict performance on numerous external criteria. At this point the issue is no longer test construction (test length) but test scoring. The usual practice is to average all trials given. If, however, a forward optimum in predictive validity exists, then averaging only those trials up to and including the optimum will yield a higher predictive validity than the usual practice. Since the differential composition of a test may change with practice and an external criterion may be most strongly related to those components of a test that predominate at the beginning (say) or in the middle of a practice series, stability and validity optima do not necessarily fall on the same trial. For the same reasons, the optimal forward average for purposes of prediction may vary from one external criterion to another.

Averaging from the first trial forward is only one way to generate a series of averages from a series of test trials. Another way is to average from the last trial backwards. The temporal stability of backward averages can, of course, be calculated and sometimes a maximum occurs before the first trial is reached (a backward optimum). Backward optima, however, are not informative about how changes in test length might affect temporal stability. A forward average of, say, 5 trials retains its meaning (refers to the same trials) regardless of how many trials are ultimately given. A backward average of 5 trials, however, refers to trials 6-10 if 10 trials are given and to trials 11-15 if a total of 15 trials is given. A backward average changes its meaning when the total number of trials changes. As a consequence, no conclusions regarding changes in test length can be drawn from a backward stability optimum.

Backward averages may also be correlated with an external criterion and, when they are, the correlation (predictive validity) rises at first and may reach an optimum prior to the first trial. In these cases, as in the corresponding cases involving forward optima, averaging only those trials up to and including the optimum (following it in the practice series) yields a higher predictive validity than averaging all trials given. Backward optima are especially helpful in improving a test's validity when a forward validity optimum also exists.

One final point should be noted. It may happen that $\bar{r}_{ii}$ and, therefore, temporal stability, take different values in different subsets of trials. Where this happens, one may restructure the test to consist exclusively of subsets with high values of $\bar{r}_{ii}$, very much as in conventional item analysis. Once such a restructuring is done, however, it should be followed (a) by forward averaging to determine the optimal test length for temporal stability and (b) by both forward and backward averaging to determine optimal scoring for predictive validity.

METHODS

The Project-A Tests

Project A is a large, multi-year effort to improve the Armed Forces Vocational Aptitude Battery (Eaton, Hanser, & Shields, 1985; Peterson, 1987). Included in this effort are 10 newly developed, computer-administered performance tests. The following brief descriptions of the 10 tests are in the same order as the tests are administered. The number of trials a subject receives on each test is given in Table 3.

Simple Reaction Time. The subject is instructed to place his or her hands in the Ready position. When the word YELLOW appears in a display box,

the subject strikes the yellow key on the test panel as quickly as he or she can. The dependent measure is average time to respond.

Choice Reaction Time. This test is much the same as Simple Reaction Time. The major difference is that the stimulus in the display box is BLUE or WHITE (rather than YELLOW), and the subject is instructed to strike the corresponding blue or white key on the test panel. The dependent measure is average time to respond on trials in which the subject makes the correct response.

Short-Term Memory. A stimulus set, consisting of 1, 3, or 5 letters or symbols, is presented on the display screen. Following a delay period, the set disappears. When the probe stimulus appears, the subject must decide whether or not it was part of the stimulus set. The dependent measure is average time to respond on trials in which the subject makes the correct response.

Target Tracking 1. This is a pursuit tracking test. The subject's task is to keep a crosshair centered within a box that moves along a path consisting exclusively of vertical and horizontal lines. The dependent measure is the average distance from the crosshair to the center of the target box.

Perceptual Speed and Accuracy. This test measures a subject's ability to compare rapidly two stimuli presented simultaneously and determine whether they are the same or different. The stimuli may contain 2, 5, or 9 characters and the characters may be letters, numbers, or other symbols. The dependent measure is average time to respond on trials where the subject's response is correct.

Target Tracking 2. This test is the same as Target Tracking 1, except that the subject uses two sliding resistors instead of a joystick to control the crosshair. The dependent measure is the same as in Target Tracking 1.

Number Memory. The subject is presented with a number on the computer screen. When the subject presses a button, the number disappears and another number appears along with an operation term (e.g., "Add 9" "Multiply by 3"). When the subject presses a button, another number and operation term are presented. This procedure continues until finally a solution to the problem is presented. The subject must then indicate whether the solution presented is correct or incorrect. The dependent measure is total time to respond on trials in which the subject correctly identifies the solution presented as itself correct or incorrect.

Cannon Shoot. The subject's task is to fire a shell from a stationary cannon so that it hits a target moving across the cannon's line of fire. The dependent measure is a deviation score indicating the difference between time of fire and optimal fire time (for example, direct hit yields a deviation score of zero).

Target Identification. The subject is presented with a target and three stimulus objects. The objects are pictures of tanks, planes, or helicopters. The target is the same as one of the three stimulus objects but rotated or reduced in size. The subject must determine which of the three stimulus objects is the same as the target object. The dependent measure is average time to respond on trials in which the subject makes the correct response.

Target Shoot. The subject's task is to move a crosshair over a moving target and then press a button to fire. The dependent measure is distance from the crosshair to the center of the target when the subject fires.

The Criterion Task

In addition to the Project-A tests, each subject was administered a criterion task. This task was Anti-Aircraft, game #1 in the Atari Air-Sea Battle cartridge (CX-2624). In this game the subject controls a gun placed two thirds of the way from left to right at the bottom of the television screen. Four different kinds of aircraft traverse the screen above the gun, in different numbers, at different speeds and altitudes, and from left to right or vice versa. The purpose of the game is to shoot down as many aircraft as possible in a 2-min-and-16-sec game. The control devices are a joystick for positioning the gun and a button for firing the missile. The missile itself was the smaller of two possible sizes (difficulty position "A"). The dependent measure is number of aircraft shot down per game.

Anti-Aircraft is a complex psychomotor skill with a high ceiling. No subject comes close to reaching the maximal possible performance with the amount of testing given.

Subjects and Procedures

The subjects were 102 central Pennsylvania undergraduate college students, 50 men and 52 women. Each subject was administered the Project-A tests at the start of the fall semester (September, October) and then again four months later at the start of the spring semester (January, February). The Project-A tests were taken in a single sitting that lasted between 45 and 75 mins, depending on how quickly the subject responded to the tests and the instructions that preceded them. The entire administration, both test and retest, instructions as well as the tests themselves, was computer controlled.

In the fall, following the Project-A tests, each subject was administered five sessions of Anti-Aircraft, each session consisting of seven games or a

little more than 16 mins of playing time. All five sessions were completed within a ten-day period, with no more than two sessions taking place on a given day. In the spring semester, again following the Project-A tests, each subject was given three sessions of Anti-Aircraft with the same number of games per session and the same conditions as to distribution as in acquisition.

RESULTS AND DISCUSSION

Table 2 presents the temporal stabilities and predictive validities of the 10 Project-A tests. "Temporal stability" in Table 2 refers to the correlation between the average of all trials given at test and at retest.

Table 2. Four-month temporal stability and predictive validity for Anti-Aircraft of the Project-A, computer-administered tests.

Test	Temporal Stability	Predictive Validity
Simple Reaction Time	.505	.251
Choice Reaction Time	.767	.117
Short-Term Memory	.694	.237
Target Tracking 1	.894	.696
Perceptual S A	.726	.107
Target Tracking 2	.910	.696
Number Memory	.689	.333
Cannon Shoot	.534	.594
Target ID	.710	.196
Target Shoot	.710	.510

"Predictive validity" refers to the correlation between the average of all trials at test and the average score per game on Anti-Aircraft in the first retest session.

In general, the 4-month temporal stabilities are better than might have been expected from one-day, test-retest reliabilities with similar tests (Kyllonen, 1985). The stabilities for Simple Reaction Time and Cannon Shoot (.505 and .534) are too low. On the other hand, those for Target Tracking 1 and 2 (.894 and .910) are excellent. The remaining six tests all have stabilities in the neighborhood of .70. The predictive validities for the two tracking tests (both .696) are outstanding and those for the two shooting tests (.595 and .510) excellent; the other six tests have low predictive validities. Note that the predictive validity of Cannon Shoot is higher than its temporal stability over the same period of time.

Three of the ten tests (Choice Reaction, Target Tracking 2, and Cannon Shoot) show a forward stability optimum. The optimum stabilities, as can be seen in Figure 1, are very little larger than what obtains when all trials given are averaged, although the drop for Cannon Shoot from Trial 35 to Trial 36 is significant at the .06 level, one-tailed. The drop yields a z-score of 1.62 by the standard test for a difference between two correlations not sharing a common variable and based on the same subjects (Steiger, 1980, p. 247, Equation 15). The issue with respect to these optima is their existence or nonexistence and not their precise location or how steeply the stability curve falls away after an optimum is reached. If the optima are real (do not result from capitalization on chance), it would make little difference if they lay a few trials distant from where they appear in Figure 1 or if points following an optimum were only a trifle lower than it. The implications for

lengthening or shortening the tests would be much the same. The stability of the three tests could not be improved by lengthening them and, if the optima remained where they now are, Choice Reaction and Target Tracking 2 might even be shortened a bit with no loss of temporal stability.

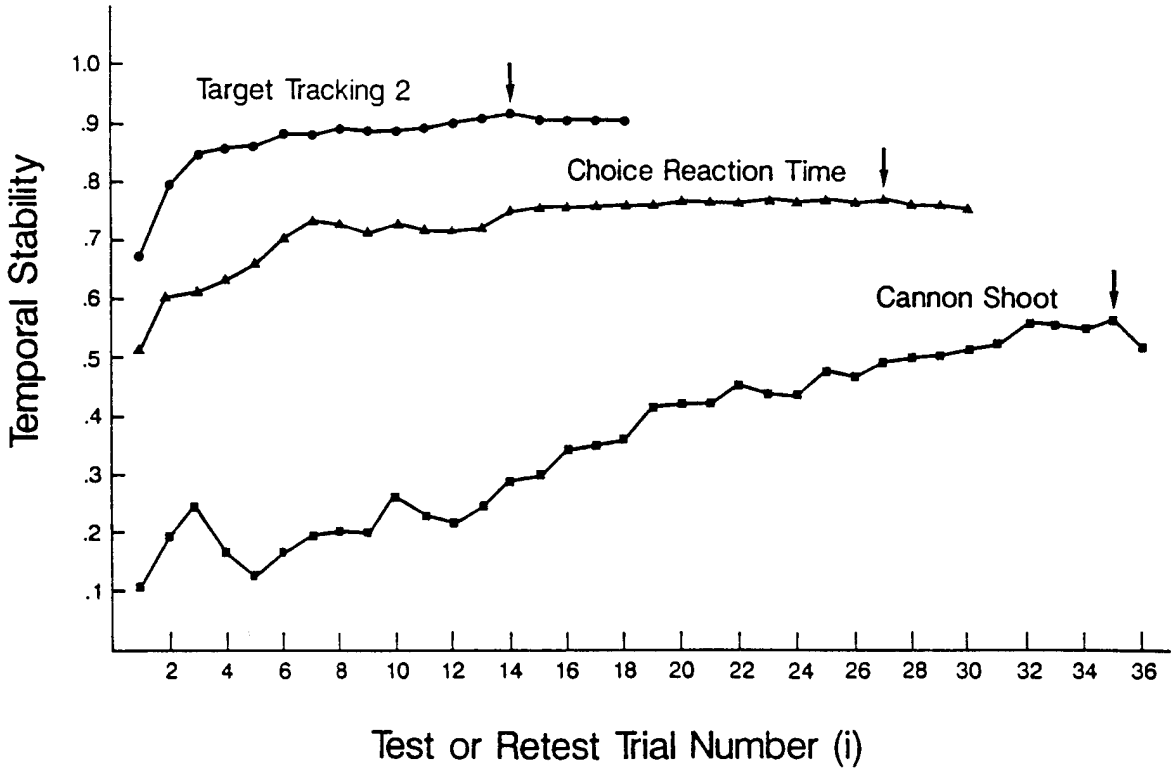

Figure 1. Four-month temporal stability of forward averages ($\bar{X}_i, i \leq n$) for Choice Reaction Time (n=30), Target Tracking 2 (n=18), and Cannon Shoot (n=36), where n indicates the total number of trials given.

Cannon Shoot, however, is the most interesting of the three tests, because in its case temporal stability appears to limit predictive validity for Anti-Aircraft.* As already noted, the predictive validity of Cannon Shoot is higher than its stability. In its case, therefore, one might expect the forward stability and validity curves to follow similar courses--and so they do. For two of the three Anti-Aircraft retest sessions Cannon Shoot shows a forward validity optimum and in both cases at Trial 35. In this connection, it should be noted that the drop in stability at Trial 36 could be an "end effect" only in the Project-A retest sessions. The first time the subjects take the Project-A tests they have no idea how long the tests are going to last. Hence the validity optima at Trial 35 are not due to an end effect (because the Project-A retest sessions are not involved). These considerations do not, of course, obviate the need for cross-validation, but just the contrary.

If, however, there really is a forward stability optimum at Trial 35 in Cannon Shoot, the fact would force a general restructuring of the test. If the temporal stability of Cannon Shoot could be increased, its predictive validity for Anti-Aircraft might exceed that of the two tracking tests. An optimum at Trial 35, however, precludes any such increase by increasing test length. If, therefore, the optimum could be confirmed, other strategies would have to be invoked. One could, for example, try giving the test in two

---

*Validity may, of course, be higher than temporal stability, if the criteria is more reliable (stable) than the predictor. Nevertheless, stability lower than a test's validity may reasonably be said to limit it.

bouts of 18 or more trials each rather than a single bout of 36 trials. One could also look for subsets of trials on Cannon Shoot with high test-retest correlations and restructure the test to consist exclusively of such subsets.

Six of the ten tests show forward validity optima for the first retest session on Anti-Aircraft. Figure 2 presents results for three of these tests. Seven tests show backward validity optima for the same criterion, but the gains over conventional scoring are smaller than for the forward validity optima.

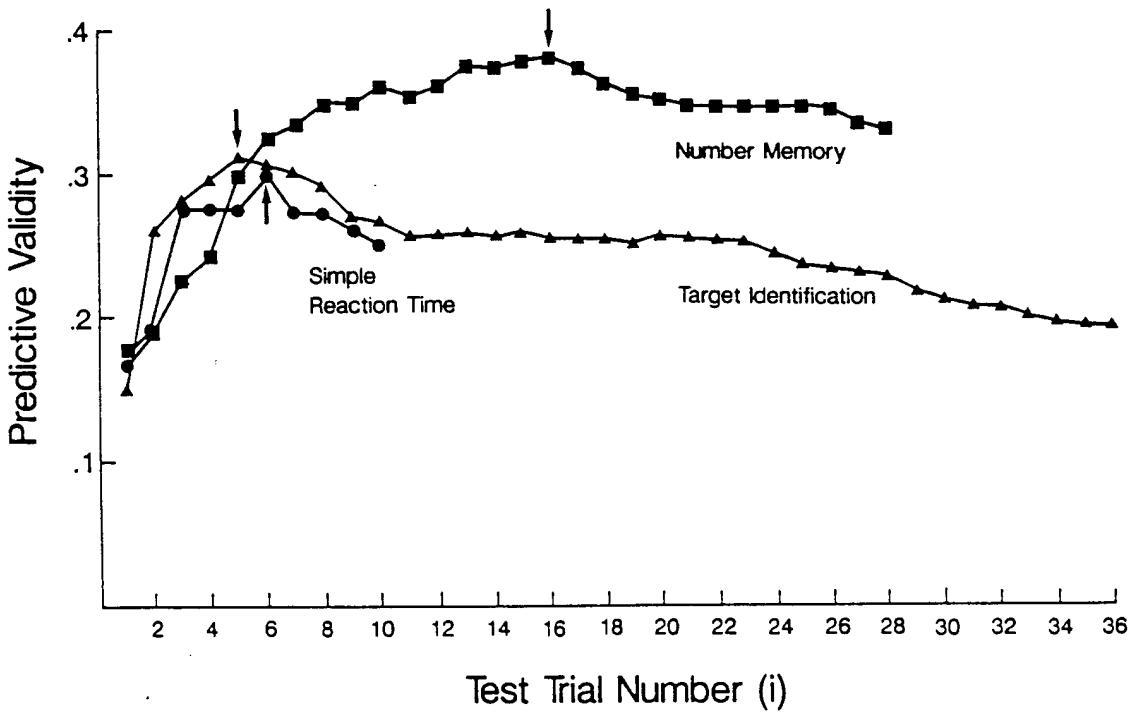

Figure 2. Predictive validity for Anti-Aircraft of forward averages ($\bar{X}_i$, $i \leq n$) for Simple Reaction Time (n=10), Number Memory (n=28), and Target Identification (n=36), where n indicates the total number of trials given.

By "optimal averaging" I mean finding that series of consecutive trials which yields the best predictive validity or an acceptably large one. The restriction to consecutive trials is helpful in reducing the total number of distinct series that must be examined before an optimal average can be determined. Unfortunately, it still leaves a large number. N test trials generate N(N+1)/2 distinct series of consecutive trials. If N equals 36, that makes 666 series and the probability of obtaining an average that is "optimal" mainly because of upward chance variations becomes substantial.

The most straightforward way to avoid excessive capitalization on chance is to limit the number of series that one examines. Accordingly, in defining an optimal validity average I have adopted the following three-step algorithm:

1, If neither a forward nor a backward optimum exists, then the optimal average is the average of all trials given (the conventional average).

2, If a forward optimum exists but not a backward optimum, the optimal average is the average of all trials from the first up to and including the optimal trial. Similarly, if a backward optimum exists but not a forward optimum, the optimal average is the average of all trials from the last back to the optimal trial.

3, If both forward and backward optima exist, the average of all trials spanned by the two optima is usually more valid than either the forward or backward optimum. If so, the optimal average is the spanning average. If not, the optimal average is the more valid of the forward and backward optima.

This algorithm requires one to examine at most 2N series. Capitalization on chance is still involved, of course, but its extent is severely controlled.

Table 3 presents the optimal averages, so defined, for the Project-A tests in predicting performance in the first retest session on Anti-Aircraft. The first column contains the total number of trials given and the next two the starting and ending trials of the optimal average. The fourth column gives the validity of the optimal average and the next column the validity of the conventional average. The next column, the sixth, presents the

Table 3. Optimal averages for the Project-A tests in predicting performance in the first retest session on Anti-Aircraft.

Test	No. of Trials	Optimal Average		Predictive Validity				
		Start	End	Opt.	Conv.	Δ	z	p
Simple Reaction	10	1	6	.299	.251	.048	1.47	<.08
Choice Reaction	30	28	30	.162	.117	.045	0.84	n.s.
Short-Term Memory	36	6	19	.291	.237	.054	2.06	<.02
Target Tracking 1	18	3	18	.698	.696	.002	0.62	n.s.
Perceptual S A	36	1	9	.222	.107	.115	1.52	<.07
Target Tracking 2	18	2	8	.717	.096	.021	1.43	<.08
Number Memory	28	4	16	.406	.333	.073	2.05	<.03
Cannon Shoot	36	10	36	.612	.594	.018	0.59	n.s
Target ID	36	1	5	.306	.196	.110	1.94	<.03
Target Shoot	30	5	29	.521	.510	.011	0.52	n.s.

difference (Δ) between the optimal and conventional validities. The right-most column but one presents the z-score (unit normal deviate) for the difference between two correlations sharing a common variable and based on the same subjects (Steiger, 1980, p. 247, Equation 14). The last column is the one-tailed significance level.

Six of the tests show differences significant at the .08 level or better. In these six tests the z-scores range from 1.43 to 2.06 and the gains obtainable by optimal averaging from .021 to .115.

These results are for single tests. We may also ask how much difference optimal averaging makes in the validity of best composites of the Project-A tests. The validity of the six tests other than the two tracking and the two shooting tests are representative of real-world, job-performance validities (Ghiselli, 1966; Schmidt, Hunter, & Pearlman, 1981). If these six tests are scored in the usual way, they yield a multiple correlation of .413. If the same tests are scored by optimal averages, the multiple correlation is .496. Adjusted for shrinkage, the correlations are .344 and .446 for conventional and optimal scoring respectively. The differences between the two multiple correlations (not adjusted) yields a z-score of 2.03, significant at the .03 level. In short, for tests with representative validities optimal scoring may improve the validity of a test composite by as much as .10.

If composites are formed of all ten tests, the gains obtainable by optimal averaging are very much less, less than .02. This result is not happenstance. First, tests with high validity (>.5) do not benefit much from optimal averaging. The point is best appreciated by considering a single

predictor in relation to two criteria, one which it predicts well and the other poorly. The validity of the test when it is i units long,

$$R_{ix} = \bar{r}_{ix}\sqrt{R_i/\bar{r}_i} \quad .$$

Increasing i increases the root-ratio by the same amount for both criteria. Hence, it takes a larger drop in the average trial validity up to trial i, $\bar{r}_{ix}$, to overcome that increase (produce an optimum) when $\bar{r}_{ix}$ is high than when it is low. For example, if the root-ratio increases from 1.00 to 1.10 as test length increases, then R_{ix} increases from .70 to .77 if $\bar{r}_{ix}$ = .70 but only from .30 to .33 if $\bar{r}_{ix}$ = .30. A drop in $\bar{r}_{ix}$ sufficient to produce an optimum has to be roughly twice as large in the former as in the latter case. In terms of variance, or z-transform, the difference between the two cases is even larger. This argument is weakened but not nullified by the tendency for tests with high validities to have higher than average reliabilities.

Second, the gains obtainable by optimal averaging in individual tests do not communicate themselves to composites which include highly predictive tests. The six tests exclusive of the tracking and shooting tests account for roughly 20% of the variance in Anti-Aircraft. The same six tests, however, account for less than 5% of the variance in Anti-Aircraft additional to that accounted for by the tracking and shooting tests. Optimal averaging increases the variance accounted for by the six tests by roughly the same amount ($\simeq$35%) whether the four "big" predictors are included or not. But the absolute amount of variance accounted for by the six "lesser" predictors

drops by a factor of four when the four "big" predictors are included. Hence, the absolute difference that optimal averaging makes drops by a factor of approximately four.

These considerations are general. There is good reason to expect that optimal averaging will improve the validities of tests and test composites only in the range from .00 to .50. This, however, is precisely where the vast majority of predictive validities lie. In practical terms, gains in predictive validity on the order of .05 to .10 in tests designed to be used on a mass basis for personnel assignment are important. These gains, moreover, are a matter of scoring only and are obtainable at no cost in test modification or testing time.

As a general approach to the construction and validation of performance tests, optimal averaging depends mainly on analytic considerations and two major empirical results: the superdiagonal patterning of intertrial correlations and the tendency for a task's differential composition to change with practice. The particular findings reported in this paper are, however, another matter. Even an algorithm as severely restricted as the one used to define an "optimal average" capitalizes to some extent on favorable chance variations. The optimal averages reported in Table 3 should, therefore, be cross-validated.

CONCLUSIONS AND MILITARY APPLICATIONS

Any test that requires a subject to demonstrate what he or she can do (rather than what he or she knows) qualifies as a performance test. Therefore, tests designed to assess information-processing parameters are performance tests. This class of tests is relevant to a broad range of military occupational specialties, ranging from gunner or pilot to radar

operator or typist. The advent of microcomputer technology has made performance testing much more feasible than ever before; and the armed services have been quick to recognize the fact, as the inclusion of ten computer-administered performance tests in the Project-A battery makes clear. Optimal averaging is an approach to the construction and validation of performance tests that recognizes and capitalizes upon their distinctive properties. In the present study, optimal averaging improved the four-month predictive validity of the Project-A tests by amounts ranging from .02 to .12. Improvements in this range have practical importance and can, moreover, be realized "for nothing." It is as easy to score a test for the optimal average as for all trials given.

Finally, there is nothing special about the criterion (Anti-Aircraft) used in this study. If optimal averaging can improve the quality of the Project-A tests for Anti-Aircraft, there is every reason to believe that it can also improve the validity of the same tests for field criteria. A general improvement, however, on the order of .05 to .10 in the predictive validity of the Project-A, computer-administered tests would be a significant advance in applied psychological testing and have sizable economic benefits for the Army, as for the other armed services.

REFERENCES

Ackerman, P.L. (1987). Individual differences in skill learning: an integration of psychiatric and information processing perspectives. Psychological Bulletin, 102, 3-27.

Bittner, A.C., Jr., Carter, R.C., Krause, M., & Harbeson, M.M. (1983). Performance Evaluation Tests for Environmental Research (PETER): Moran and computer batteries. Aviation, Space, and Environmental Medicine, 54, 923-928.

Eaton, N.K. & Shields, J. (1985). Validating selection tests for job performance. In J. Zeidner (Ed.), Human productivity enhancement. Vol. 2, Acquisition and development of personnel. New York: Praeger.

Fleishman, E.A. & Hempel, W.E., Jr. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 19, 239-252.

Ghiselli, E.E. (1966). The validity of occupational aptitude tests. New York: Wiley.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley & Sons.

Humphreys, L.G. (1960). Investigation of the simplex. Psychometrika, 20, 173-192.

Jones, M.B. (1962). Practice as a process of simplication. Psychological Review, 69, 274-294.

Jones, M.B. (1969). Differential processes in acquisition. In E.A. Bilodean (Ed.), Principles of skill acquisition. New York: Academic Press, 1969.

Jones, M.B. (1989a). Individual differences in skill retention. American Journal of Psychology, 102, 183-196.

Jones, M.B. (1989b). Slope-controlled performance testing (Final Report on Grant No. AFOSR-87-0216A). Washington, DC: Air Force Office of Scientific Research.

Jones, M.B. (1989c). Optimal averaging in performance tests (Final Report on Contract No. MDA 903-86-C-0145). Alexandria, VA: U.S. Army Research Institute.

Kennedy, R.S., Bittner, A.C., Jr., Carter, R.C., Krause, M., Harbeson, M.M., McCafferty, D.B., Pepper, R.L., & Wiker, S.F. (1981). Performance Evaluation Tests for Environmental Reserach (PETER): Collected papers (NBDL-80R008). New Orleans, LA: Naval Biodynamics Laboratory.

Kyllonen, P.C. (1985). Theory-based cognitive assessment (AFHRL-TP-85-30). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Melton, A.W., Ed. (1947). Apparatus tests. Washington, DC: U.S. Government Printing Office (AAF Aviation Psychology Program Research Report No. 4).

Messick, S., & Jungblut, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, 89, 191-216.

Peterson, N.G. (Ed.).(1987). Development and field test of the trial battery for Project A (Technical Report 739). Alexandria, VA: U.S. Army Research Institute.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: a red herring. Journal of Applied Psychology 66, 166-185.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Wing, H. (1980). Practice effects with traditional test items. Applied Psychological Measurement, 4, 141-155.