Technical Report 600

(12)

Military Experts' Estimates of Continuous Operations Performance (or Close but No Cigar)

Karen L. Neff and Robert E. Solick

ARI Field Unit at Fort Leavenworth, Kansas
Systems Research Laboratory

U. S. Army
Research Institute for the Behavioral and Social Sciences

November 1983

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
Technical Director

L. NEALE COSBY
Colonel, IN
Commander

Technical review by

Ira T. Kaplan
Helen V. Lewis

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

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 findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report 600	2. GOVT ACCESSION NO. AD-A140220	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Military Experts' Estimates of Continuous Operations Performance (Or Close But No Cigar)		5. TYPE OF REPORT & PERIOD COVERED -- 6. PERFORMING ORG. REPORT NUMBER --
7. AUTHOR(s) Karen L. Neff Robert E. Solick		8. CONTRACT OR GRANT NUMBER(s) --
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Army Research Institute 5001 Eisenhower Avenue Alexandria, Virginia 22333		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q162717A790
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute 5001 Eisenhower Avenue Alexandria, VA 22333		12. REPORT DATE November 1983 13. NUMBER OF PAGES 50
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

continuous operations performance
performance under stress
performance prediction
ENDURE
FDC
Early Call I
Early Call II
expert's estimates
performance models

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The feasibility of supplementing human performance data with estimates of performance in adverse environments was examined for cases where no hard data is available for use in land combat models. The accuracy of military experts' estimates of performance in continuous operations was evaluated by examining the amount of convergence between samples of estimates made by military officers and actual performance values obtained in four field exercises. There was strong agreement among the officers in their predictions of performance.

However, the officers' predictions of performance did not agree with actual performance measures obtained in the field exercises.

Technical Report 600

Military Experts' Estimates of Continuous Operations Performance (or Close but No Cigar)

Karen L. Neff and Robert E. Solick

Submitted by
Robert S. Andrews, Chief
ARI Field Unit at Fort Leavenworth, Kansas

Approved as technically adequate
and submitted for publication by
Jerrold M. Levine, Director
Systems Research Laboratory

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

November 1983

Army Project Number 2Q162717A790	Human Performance Effectiveness & Training

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

The Fort Leavenworth Field Unit of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) conducts a research program in support of the Combined Arms Center, which includes the Combined Arms Combat Developments Activity (CACDA) and the Command and General Staff College (CGSC).

The Field Unit is presently involved in assisting the local combat modeling community with the representation of human performance variables in land combat models. In the absence of performance measurements under realistic conditions, combat modelers often resort to the use of panels of military experts to provide estimates of how performance would be affected by various situational factors. The present investigation explored the validity of such judgments by asking for estimates of performance on military tasks in situations where data from controlled field exercises exist. This investigation is responsive to the objectives of Army Project 2Q162717A790 concerned with the improvement of command and control procedures and systems.

Edgar M Johnson

EDGAR M. JOHNSON
Technical Director

Accession For
NTIS GRA&I ☒
DTIC TAB □
Unannounced □
Justification________
____
By________
Distribution/
Availability Codes

Dist	Avail and/or Special
A-1	

MILITARY EXPERTS' ESTIMATES OF CONTINUOUS OPERATIONS PERFORMANCE (OR CLOSE BUT NO CIGAR)

EXECUTIVE SUMMARY

Objective:

To determine the feasibility of supplementing human performance data used in land combat models with estimates of soldier performance in adverse environments.

Procedure:

Estimates of specific performance were compared with actual performance data from previous studies. Three comparison studies were selected: (a) ENDURE, where tank crews performed simulated combat tasks over a 48-hour period, (b) a laboratory investigation in which fire direction center (FDC) teams underwent up to 48 hours of simulated sustained combat operations, and (c) Early Call I and Early Call II, where parachute platoons performed a sustained tactical defensive exercise in Great Britain for up to five days without sleep. Detailed descriptions of the performance tests along with average scores or times for the first time period were given to students from the Combined Arms and Services Staff School (CAS^3). The CAS^3 students estimated the scores for the second, third, and fourth periods.

Principal Findings:

1. The officers agreed strongly among themselves in their predictions of performance. This was shown by extremely high intraclass correlations.

2. The officers' predictions of performance did not reflect actual performance measures obtained in the original field exercises. The expert raters' predictions of performance were significantly different from the actual performance measures.

3. Expert raters' estimates were no more accurate for performance after 12 hours than after 24, 36, or 48 hours of continuous operations.

4. Expert raters' predictions of performance were more accurate for cognitive and vigilance tasks than for simple motor tasks.

Utilization of Findings:

In situations similar to those described here, attempts to supplement human performance data with military experts' estimates of performance in adverse environments may result in inaccurate and possibly misleading models.

Caution is necessary in using expert ratings, even in cases of strong agreement among the raters.

MILITARY EXPERTS' ESTIMATES OF CONTINUOUS OPERATIONS PERFORMANCE

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

Military Experts' Estimates of Continuous Operations Performance (or Close but no Cigar)

Karen L. Neff
and
Robert E. Solick

INTRODUCTION

Attempts to calibrate or to assess the validity of expert judgments have led to little conclusive evidence for experts' abilities to make predictions of random events, except for meteorologists who are surprisingly good at making familiar, common-place types of forecasts (Lichtenstein, Fischhoff, and Phillips, 1977; Murphy and Winkler, 1977). Predictions for securities and the stock market generally are little better than a simple random model or a no-information strategy (Borcherding, 1978). Estimates regarding other events and other types of estimates generally lie somewhere between these two extremes. In general most types of estimates and predictions have been shown to be subject to numerous types of biases on the part of the expert which play havoc with his estimates. (Bowonder, 1981; Einhorn and Hogarth, 1981; Lichtenstein, Fischhoff, and Phillips, 1977; Slovic, Fischhoff and Lichtenstein, 1979a; Tversky, 1969; Tversky and Kahneman, 1974). There is evidence that these biases can be reduced through training, particularly when there are short, clear links between the action and its outcome (Einhorn and Hogarth, 1978; Hogarth and Makridakis, 1981); the excellent predictions made by weather forecasters are evidence for such an effect. There is also evidence that the context and form of the instigation for the prediction can have a significant effect upon the decision made (Kahneman and Tversky, 1981, 1982a, 1982b).

It is fairly common practice within the military to use experts' estimates as shortcuts in decision making or as predictions of later performance (Uhlaner and Drucker, 1980). However, few attempts have been made to determine the accuracy of the estimates or decisions. Harman and Press (1975) provide guidelines for collecting and analyzing judgments from groups of experts. They also provide recommendations for selecting a panel of experts. They note that the ideal method to assess the validity of predictions is "to compare them with actual outcomes" (p.10). However, there are difficulties in implementing this approach, particularly when one is attempting to make forecasts in the first place. Harman and Press (1975) recommend the use of a pilot study to establish the validity of the predictions wherever possible.

Ryan-Jones (1979) did attempt to evaluate the validity of military experts' judgments. His was the only such attempt available. He compared opinions of squad leaders and platoon leaders regarding task difficulty against the percentage of soldiers failing a criterion-referenced test on the same tasks. He found a non-significant correlation between the expert ratings and the independent measure of task difficulty. There was an apparent, but not statistically significant, trend toward rating difficult tasks as easy.

The purpose of the research described here was to provide additional evidence regarding the accuracy of military experts' predictions. We hoped to determine the feasibility of supplementing human performance data used in land combat models with estimates of performance in adverse environments when no hard data is available.

Background

It is widely recognized that the performance of soldiers is a prime determinant of the effectiveness of weapons, units,and forces in battle. Yet, land combat models only recently have come to reflect human performance variations and limitations. Factors relating to human performance in combat include the state of the soldier (training, morale, fatigue, fear), the state of the environment (precipitation, temperature, visibility) and the quality of command and control (as reflected in planning, decision making, intelligence gathering and communicating, as well as more charismatic aspects of leadership). These variables must be shown to influence battle outcomes through their relations to traditional model constructs such as the probability of a hit, the vulnerability of a unit or system, the likelihood that a system will be in good repair (thus able to participate in the battle), the probability that orders will be received, and the time required for functions like movement and construction of defenses. In almost all cases, the relationships between the human factors and these model constructs are not known quantitatively, although the general direction and probable magnitude of the relationships can sometimes be deduced.

For many human variables of interest, such as performance under extreme stress, it is virtually impossible to gather data on tactical performance under realistic conditions. For others, such as fatigue or level of training, it is feasible, but expensive and time consuming, to gather this data. Since the construction of crew or operator models to relate the physiological and psychological state of the soldier to system performance measures used in combat models requires such data on every task, the model developer may resort to the use of quantitative estimates of performance as a surrogate for performance data.

The accuracy of such estimates is of primary importance. Therefore, it was proposed to determine whether military experts can make quantitative estimates of sufficient accuracy for formulation of functions for incorporation in the models.

The initial effort reported here evaluated the accuracy of military experts' performance estimates by examining the amount of convergence between samples of estimates and the performance values obtained in field exercises, and by examining the amount of agreement among personnel familiar with tactical tasks concerning variations in human performance.

METHOD

Numerical estimates were gathered from a sample of Army officers, 29 students from the Combined Arms and Services Staff School (CAS^3), Fort Leavenworth. Nine had armor experience, nine had artillery experience, and eleven had infantry experience. All had at least one year of experience at the company command level.

Three different field exercises were selected for use as comparison data for experts' predictions of troop performance: (1) work unit ENDURE, in which tank crews were required to perform simulated tasks over a 48-hour period (Ainsworth & Bishop, 1971); (2) a laboratory study in which Fire Direction Center (FDC) teams underwent up to 48 hours of simulated, sustained combat operations (Banderet & Stokes, 1980; Banderet, Stokes, Francesconi, Kowal, & Naitoh, 1980); and (3) field exercises conducted in Great Britain, Exercise Early Call I and Exercise Early Call II, where parachute platoons performed a sustained tactical defensive exercise for nine days (Haslam, 1978, 1980, 1981, 1982; Haslam, Allnutt, Worsley, Dunn, Abraham, Few, Lubuc, & Lawrence, 1977). These three exercises were selected because they were the only ones available that measured sustained performance on military tasks while attempting to hold constant other factors which might affect performance.

Three questionnaires were developed based upon these three exercises. Expert raters for each questionnaire type had experience in the questionnaire's specific type of activity. Each questionnaire provided a detailed description of the context in which performance tests were conducted. Information regarding the test conditions, the experimental procedures, the type of personnel who participated, and the time schedule was outlined. Each of the three types of questionnaires contained descriptions of particular tests administered to the troops participating in the continuous operations exercises. The descriptions included how each test was scored or timed, and the average score and time obtained for the first period. After each description of a task, the participants were asked to estimate the average score and average time obtained by the soldiers for the second, third, and fourth time periods. Where there was a maximum score which could be obtained for perfect performance, the maximum score was given as an additional anchor.

Questionnaires were administered in group sessions by CAS^3 section leaders. Participants were briefed on the potential of the research for applications in modeling and doctrine. They were asked to complete the questionnaire on their own without conferring with other persons or documents. The participants were instructed to base their responses upon their own knowledge, experience, and training.

RESULTS

Figures 1 through 40 in Appendix A show averages of rater estimates of performance and field exercise measures of performance as functions of time for each of the tests from the armor and from the artillery continuous operations field exercises. Figures for all of the tests from Early Call I and Early Call II were not included because some of the information regarding the field exercises was sensitive; only test exercises which have appeared in the open literature were included as figures. These figures show raters with armor experience overestimated actual field-exercise performance data in three cases (Figures 1, 2, and 7), underestimated performance data in eight cases (Figures 4, 6, 10, 11, 12, 13, 14, and 15), made both overestimates and underestimates depending upon the time period, in four cases (Figures 3, 5, 16, and 17) and made fairly accurate predictions across all time periods in one case (Figure 9). Figures 18 through 33 show raters with artillery experience in general overestimated actual performance data in three cases (Figures 18, 24, and 33), underestimated actual performance data in eight cases (Figures 22, 23, 26, 27, 28, 29, 30, and 31) and both overestimated and underestimated actual performance data, depending upon the successive time period, in five cases (Figures 19, 20, 21, 25, and 32). Figures 34 through 40 show that for Early Call I and II raters consistently overestimated the decrement in performance (underestimated actual performance data) with successive time periods in all cases but one (Figure 40).

Inter-rater agreement for the performance ratings was estimated by the intraclass correlation for each performance measure in each task in each exercise (see Table 1). Nunnally (1967) contended that reliabilities of .60 or .50 will suffice for exploratory research. Table 1 shows acceptable levels of interrater reliability for nearly all the tasks. Sufficient agreement among the raters with armor experience for project ENDURE was found for all but five of the estimates of performance measures: (a) ditch crossing time, (b) log crossing time, (c) firing accuracy of the main gun on a stationary target with the tank stationary, (d) firing accuracy of the main gun on a moving target with the tank stationary, and (e) accuracy of completion of the maintenance checklist. For the artillery tasks, sufficient agreement was found for all of the tasks, except preplanning errors that exceeded 990 mils. Very high interrater agreement was found among officers with infantry experience. All predictions for Early Call I and Early Call II had intraclass correlations of nearly .80 or greater.

Since the response scales in the field and laboratory exercises varied from task to task, the estimate scales also varied according to task. Either the response measure had to be treated as a separate dependent variable for any parametric analysis of the responses, or the response scales had to be converted to a common measurement scale. Estimates for each performance measure were treated as separate dependent variables for the intraclass correlation computations.

Confidence intervals based upon the _t_-distribution were constructed to determine whether the estimated values were significantly different from the actual performance values obtained in the field exercises. Computations

Table 1

Reliability of Performance Estimates
by the Intraclass Correlation

Task	Intraclass correlation[a]
Armor	
Driving exercises	
Minefield-time	.84
Minefield-accuracy	.71
Slalom-time	.68
Slalom-accuracy	.76
Ditch-time	.44
Ditch-accuracy	.72
Log-time	**[b]
Log-accuracy	.72
Gunnery exercises	
Main gun-stationary tank, stationary target-time	.52
Main gun-stationary tank, stationary target-accuracy	.15
Main gun-stationary tank moving target-time	.87
Main gun-stationary tank moving target-accuracy	.47
Caliber .50 machinegun-stationary tank, moving target-time	.79
Caliber .50 machinegun-stationary tank, moving target-accuracy	.86
Coaxial machinegun-moving tank, stationary target-accuracy	.54
Maintenance-time	.80
Maintenance-accuracy	**[b]
Artillery	
Prioritizing-number	.71
Prioritizing-latency	.80
Unplanned mission errors	
Greater than 990 mils	.83
90-990 mils	.95
30-89 mils	.96
15-29 mils	.97
7-14 mils	.93

Table 1 (Continued)

Reliability of Performance Estimates
by the Intraclass Correlation

Task	Intraclass correlation
Artillery (Continued)	
Unplanned missions computation-latency	.94
Preplanning latency	.90
Preplanning-number	.60
Preplanning errors	
Greater than 990 mils	**[b]
90-990 mils	.89
30-89 mils	.82
15-29 mils	.87
7-14 mils	.87
On-call mission response latency	.97
Early Call I	
Grouping capacity	.86
Marching	.98
Weapon-handling tests	
Time to fill magazine by hand	.91
Time to load rifle, standing	.94
Time to unload rifle, standing	.93
Time to strip rifle to firing pin	.92
Time to assemble rifle	.94
Average score-strip and reassemble	.81
Vigilance shooting	.97
Commander ratings of military effectiveness	.99
Hours to withdraw	.98

Table 1 (Continued)

Reliability of Performance Estimates
by the Intraclass Correlation

Task	Intraclass correlation
Early Call II	
Vigilance shooting	.96
Vigilance with night sight	
Percentage detected	.95
False alarms	.95
Percentage detected-teams	.94
False alarms-teams	.94
Moving target shooting	
Hits	.97
Shots leading	.78
Shots lagging	.93
Grouping capacity	.92

[a] Intraclass correlations of ,5 or greater are considered acceptable levels of inter-rater reliability (Nunnally, 1967).

[b] Intraclass correlations were computed using Ebel's formula, Negative values result for F ratios less than one. They do not connote an inverse relationship, but should be considered equal to zero for purposes of interpretation; they are indicated by double asterisks,

treating each performance measure as a separate dependent variable would have resulted in a prohibitively large number of tests. Therefore, to create a common scale for all the measures, performance predictions were converted to ratios by dividing the rater estimate by the performance score obtained in the field exercise. Log transforms of the ratios eliminated their positive skew (Nickerson, 1981). If the rater estimate was greater than the actual performance score, the transformed value was positive. If the estimate was less than the actual performance score, the transformed value was negative. The log transform of the ratio was zero if the predicted score was equal to the actual performance score.

Confidence intervals based upon the _t_-distribution were constructed for the log transforms of the ratios of predicted to actual performance. Confidence intervals which included zero (the $\log_{10}$ of 1) as a possible mean would indicate that the predicted values were not significantly different from the actual performance values obtained in the field exercises. As can be seen in Table 2, none of the confidence intervals included zero. Therefore, the expert raters' predictions were significantly different from the actual performance measures for soldiers participating in the four field exercises.

Three performance measures on the questionnaires and included in the intraclass correlation computations were not considered in any further statistical analyses. On-call mission response latency predictions made by artillerymen were excluded because 30 was inadvertently given as the anchor for their predictions, rather than 10, which was the correct average response latency for on-call missions in the field exercises. Even through given an anchor inflated by 300%, the artillerymen's estimates of that item had an intraclass correlation of .97. Two measures from Early Call I, commander rating of military effectiveness and observer ratings of marching performance, were excluded because their scale of measurement did not justify conversion to ratios.

Results of ANOVAs for the four field exercise show significant main effects for tasks for armor ($F = 5.5936$; $df = 16.32$; $p < .01$), artillery ($F = 11.8140$; $df = 14.28$; $p < .01$), and Early Call I ($F = 8.4367$; $df = 7.21$; $p < .01$). Main effects for tasks were not significant for Early Call II. Main effects for fatigue were not significant for any of the four questionnaire types. Simple effects were explored using Tukey's HSD (honesty significant difference test) to make pairwise comparisons between means.

Statistically significant differences in the officers' ability to predict performance for armor groups were found between the following tasks: (a) minefield accuracy and slalom accuracy, (b) minefield accuracy and ditch accuracy, (c) minefield accuracy and log accuracy, (d) minefield accuracy and main gun time with stationary tank and stationary target, (e) minefield accuracy and caliber .50 machinegun accuracy with stationary tank and moving target, (f) log accuracy and minefield time, (g) log accuracy and slalom time, (h) log accuracy and main gun accuracy with stationary tank and stationary target, (i) log accuracy and main gun time with stationary tank and moving target, (j) log accuracy and caliber .50 machinegun time with stationary tank and moving target, (k) log accuracy and coaxial machinegun

Table 2

Means, Standard Error of the Means and Confidence Intervals for Log Transforms of Ratios of Predicted to Actual Performance

	Mean	Standard Error of the Mean	*df*	*t* Distribution Confidence Interval
Armor	.0568	.0186	49	C.99 $\left\{.0114 \leq \bar{X} \leq .1022\right\}$
Artillery	.0290	.0200	43	C.99 $\left\{.0251 \leq \bar{X} \leq .0831\right\}$
Early Call I	.1310	.0362	30	C.99 $\left\{.0314 \leq \bar{X} \leq .2306\right\}$
Early Call II	.0519	.0474	25	C.99 $\left\{.0802 \leq \bar{X} \leq .1840\right\}$

Note: The means of the distributions of ratios are zero (the $\log_{10}$ of 1), if the predicted values are equal to the actual performance scores.

accuracy with moving tank and stationary target, (l) caliber .50 machinegun accuracy with stationary tank and moving target and slalom accuracy, (m) caliber .50 machinegun time with stationary tank and moving target and main gun time with stationary tank and stationary target, (n) caliber .50 machinegun time with stationary tank and moving target and caliber .50 machinegun accuracy with stationary tank and moving target, (o) caliber .50 machinegun accuracy with stationary tank and moving target and slalom time, (p) caliber .50 machinegun accuracy with stationary tank and moving target and main gun accuracy with stationary tank and stationary target, (q) caliber .50 machinegun accuracy with stationary tank and moving target and main gun time with stationary tank and moving target.

Statistically significant differences in the officers' ability to predict performance for artillery teams were found between prioritizing latency and each of the other tasks: (a) prioritizing latency and prioritizing number, (b) prioritizing latency and unplanned mission errors greater than 990 mils, (c) prioritizing latency and unplanned missions computation latency, (d) prioritizing latency and unplanned mission errors from 90 to prioritizing latency and unplanned mission errors from 30 to 89 mils, (f) prioritizing latency and unplanned mission errors from 15 to prioritizing latency and unplanned mission errors from 7 to 14 mils, (h) prioritizing latency and preplanning latency, (i) prioritizing latency and preplanning number, (j) prioritizing latency and preplanning errors greater than 990 mils, (k) prioritizing latency and preplanning errors from 90 to 990 mils, (l) prioritizing latency and preplanning errors from 30 to 89 mils, (m) prioritizing latency and preplanning errors from 15 to 29 mils, and (n) prioritizing latency and preplanning errors from 7 to 14 mils.

Using Tukey's HSD test, statistically significant differences in the officers' ability to predict infantry performance in Early Call I were found between vigilance shooting and all of the other tasks, except the weapon handling average score: (a) vigilance shooting and grouping capacity, (b) vigilance shooting and time to fill magazine by hand, (c) vigilance shooting and time to load rifle, (d) vigilance shooting and time to unload rifle, (e) vigilance shooting and time to strip rifle to firing pin, and (f) vigilance shooting and time to assemble rifle.

DISCUSSION

The performance predictions were examined in terms of two questions: Did the officers agree among themselves in their predictions of performance, and did the officers' predictions of performance reflect actual performance measures obtained in the original field exercise? The answer to the first question was yes; the answer to the second question was no.

Intra-class correlations revealed high inter-rater reliabilities for most items for each questionnaire type.

Confidence intervals based upon the _t_-distribution showed that the expert raters' predictions were significantly different from the actual performance measures for soldiers participating in the four field exercises.

ANOVAs revealed no significant difference in the raters' ability to predict performance scores as a function of length of sustained performance. The raters were not significantly more accurate in their estimates of soldier performance after the soldiers had undergone 12 hours of continuous operations than after the soldiers had undergone 24 hours or 48 hours of continuous operations.

ANOVAs did reveal that the raters' predictions were significantly more accurate for some of the tasks than for others. This was true of each questionnaire type, except Early Call II where no difference was found among tasks in the accuracy of predictions.

Consistent agreement on incorrect predictions is fairly clear evidence of systematic bias.[1] To explore the nature of this bias, a post hoc analysis was performed on the performance estimates provided for exercise Early Call II. Briefly, it was hypothesized that the estimation task was too difficult to perform based upon past experience with the military tasks and that the estimators as a group adopted a simplification strategy, basing their predictions upon a simple, but inappropriate, qualitative model of the effect of fatigue on performance. Four such models were explored. The degree of fit was determined by classifying each individual's predictions for each of nine tasks in Early Call II according to whether or not they violated any of the assumptions of each model.

The least restrictive model assumed that performance would remain the same or deteriorate with increasing levels of fatigue, but that it would not get better. This model fit 85 of the 99 cases, where a case consisted of three predictions about a single task by one rater. One person accounted for 8 of the 14 deviations from the model by consistently predicting that performance would recover in the last time interval of the field experiment.

The next least restrictive model assumed strictly decreasing performance with increasing levels of fatigue. This model was consistent with 67 of the 99 cases. One additional person consistently violated this model by assuming no decrease in performance from the first to the second time interval.

Two more restrictive models were examined as approximations to the effects of fatigue on vigilance tasks, since these tasks are more prone to fatigue effects and thus more likely to be within the experience of the predictors. The partially ordered intervals model assumed that performance would get worse over successive time periods and that the amount by which it worsened would be

[1] The obvious alternative explanation is that the estimators conferred among themselves despite the instructions. Monitoring by section leaders precluded this possibility.

the same or greater over successive time periods. The most restrictive model considered was the strictly ordered intervals model, which assumed that the performance decrements due to fatigue would increase over successive time periods.

The more restrictive models did not fare as well. Partially ordered intervals fit 48 of the 99 cases. Strictly ordered intervals fit in only 30 cases. Neither model fit any individual on all tasks. Cases deviating from these models (other than the cases previously described) tended to show a floor effect, with estimated performance dropping rapidly, then more slowly approaching the minimum performance obtainable on a task.

In summary, the high correlations among estimates appeared to be due to a large proportion of the expert raters adopting similar simplification strategies. In those cases where their simple model happened to fit the situation, the group's average judgment was quite accurate. However, the predictions did not appear to take into account the differing effects of fatigue on various types of activity. The groups tended to overestimate the effects of fatigue on simple motor skills, to underestimate the effects on cognitive and perceptual tasks, and to ignore the effects of potentially confounding variables, such as learning, lighting, diurnal variations, and knowledge that the end of the task was near.

A few predictors appeared to deviate from the simplest model. One consistently assumed that performance would recover in the last time period (as it often did). One assumed that fatigue would have no effect on simple motor skills and very little effect on other tasks; he was closest to the field data on the motor skills and the least accurate predictor on the vigilance tasks.

The research reported here and earlier suggests that supposed experts in general, excepting meteorologists in some situations, often make no better predictions than simple random models are capable of making and that experts often are actually poorer predictors (Borcherding, 1978; Hogarth, and Makridakis, 1981; Lichtenstein, Fischhoff, and Phillips, 1977; Murphy, and Winkler, 1977; Slovic, Fischhoff, and Lichtenstein, 1977). Perhaps random, no-information models should be used in simulations in place of expert estimates of doubtful or unconfirmed validity. Bootstrapping, which replaces judges with algebraic models of their own weighting policies, has resulted in models that perform as well or better than the judges themselves (Slovic, Fischhoff, and Lichtenstein, 1977). Finally, Dawes and Corrigan (1974) have demonstrated the extreme robustness of the simple linear model which is able to capture most of the variance in many judgmental and decision making situations, even in instances considered to be inherently non-linear. This suggests the use of a simple model where actual performance measures are not available -- essentially the same simplification strategy that the expert raters appeared to adopt for this experiment.

CONCLUSION

Expert raters' subjective estimates of performance were obtained for tasks with known objective measures. The purpose was to establish the validity or lack of validity of such expert-rater estimates. While a high level of inter-rater reliability was found, the ratings were shown to have little or no relationship to actual performance under the described circumstances. Therefore, one must conclude that in situations similar to those described here, attempts to supplement human performance data with military experts' estimates of performance in adverse environments may result in inaccurate and possibly misleading combat models.

When raters show a high degree of inter-rater reliability, there may be a temptation to accept the ratings as accurate, when in fact the ratings may be systematically biased. The results reported here demonstrate the need for caution in accepting expert ratings, even in cases of high interrater reliability.

REFERENCES

Ainsworth, L.L. and Bishop, H.P. The effects of a 48-hour period of sustained field activity on tank crew performance. Alexandria, Virginia: Human Resources Research Institute Technical Report 71-16, July, 1971. (NTIS No. AD-731 219)

Banderet, L.E., and Stokes, J.W. Simulated, sustained-combat operations in the field artillery Fire Direction Center (FDC): A model for evaluating biomedical indices. Proceedings of the Army Science Conference, 1980, 19, 167-181.

Banderet, L.E., Stokes, J.W., Francesconi, R., Kowal, D.M., and Naitoh, P. Artillery teams in simulated sustained combat: Performance and other measures. In L.C. Johnson, D.I. Tepas, W.P. Colquhoun, and M.J. Colligan (Eds.), Biological rythms, sleep and shift work (Spectrum Pub 7). New York: Spectrum, 1981.

Borcherding, K. An attempt to use portfolio theory as a decision aid. In G.Pask (Ed.), Current approaches to decision making in complex systems: II. Alexandria, Virginia: Army Research Institute Technical Report TR-78-B4, March, 1978. (Summary)

Bowonder, B. Issues in environmental risk assessment. Journal of Environmental Systems, 1981, 10, 305-331.

Dawes, R.M., and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Einhorn, H.J., and Hogarth, R.M. Confidence in judgment: Persistence of the illusion of validity. Psychological Review, 1978, 85, 395-416.

Einhorn, H.J., and Hogarth, R.M. Behavioral decision theory: Processes of judgment and choice. Annual Review of Psychology, 1981, 32, 53-88.

Harman, A.J., and Press, S.J. Collecting and analyzing expert group judgment data. Santa Monica, California: The Rand Corporation Paper Series P-5467. July, 1975.

Haslam, D.R. The effect of continuous operations upon the military performance of the infantryman (Exercise "Early Call II"). Farnborough, Great Britain: Army Personnel Research Establishment Report 4/78, 1978. (UK Restricted)

Haslam, D.R. The effects of sleep loss upon the motivation, morale, and mood of the soldier. Paper presented at NATO symposium on Motivation and Morale in the NATO Forces, Brussels, 1980.

Haslam, D.R. The military performance of soldiers in continuous operations: Exercises "Early Call I and II. In L.C. Johnson, D.I. Tepas, W.P. Calquhoun, and M.J. Colligan (Eds.), Advances in sleep research (Vol. 7). New York: Spectrum, 1981.

Haslam, D.R. Sleep loss, recovery sleep, and military performance. Ergonomics, 1982, 25(2), 163-178.

Haslam, D.R., Allnutt, M.F., Worsley, D.E., Dunn, D., Abraham, P., Few, J., Labuc, S., and Lawrence, D.J. The effect of continuous operations upon the military performance of the infantryman (Exercise "Early Call"). Farnborough, Great Britain: Army Personnel Research Establishment Report 2/77, 1977. (UK Restricted)

Hogarth, R.M., and Makridakis, S. Forecasting and planning: An evaluation. Management Sciences, 1981, 27(2), 115-138.

Kahneman, D., and Tversky, A. On the study of statistical intuitions. Cognition, 1982, 11, 123-141. (a)

Kahneman, D., and Tversky, A. Variants of uncertainty. Cognition, 1982, 11, 143-157. (b)

Lichtenstein, S., Fischhoff, B., and Phillips, L.D. Calibration of probabilities: The state of the art. In H. Jungermann and G. deZeeuw (Eds.), Decision making and change in human affairs: Proceedings of the fifth research conference on subjective probability, utility, and decision making. Dordrecht, The Netherlands: Reidel, 1977.

Murphy, A.H., and Winkler, R.L. The use of credible intervals in temperature forecasting: Some experimental results. In H. Jungermann and G. de Zeeuw (Eds.), Decision making and change in human affairs: Proceedings of the fifth research conference on subjective probability, utility, and decision making. Dordrecht, The Netherlands: Riedel, 1977.

Nickerson, R.S. Motivated retrieved from archival memory. In H.E. Howe and J.H. Flowers (Eds.), Nebraska symposium on motivation 1980. Lincoln, Nebraska: University of Nebraska Press, 1981.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

Ryan-Jones, D.L. A comparison of expert ratings of task difficulty with an independent criterion. Alexandria, Virginia: Army Research Institute Technical Report 418, November, 1979.

Slovic, P., Fischhoff, B., and Lichtenstein, S. Rating the risks. Environment, 1979, 21(3), 14-20, 36-39. (a)

Slovic, P., Fischhoff, B., and Lichtenstein, S. Weighing the risks. Environment, 1979, 21(4), 17-20, 32-38. (b)

Tversky, A. Intransivity of preferences. Psychological Review, 1969, 76 , 31-48.

Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 185 1124-1131.

Tversky, A., and Kahneman, D. The framing of decisions and the psychology of choice. Science, 1981, 211, 453-458.

Uhlaner, J. E., and Drucker, A. J. Military research on performance criteria: A change of emphasis. Human Factors, 1980, 22(2), 131-139.

APPENDIX A PREDICTED PERFORMANCE ON FORTY MILITARY TASKS

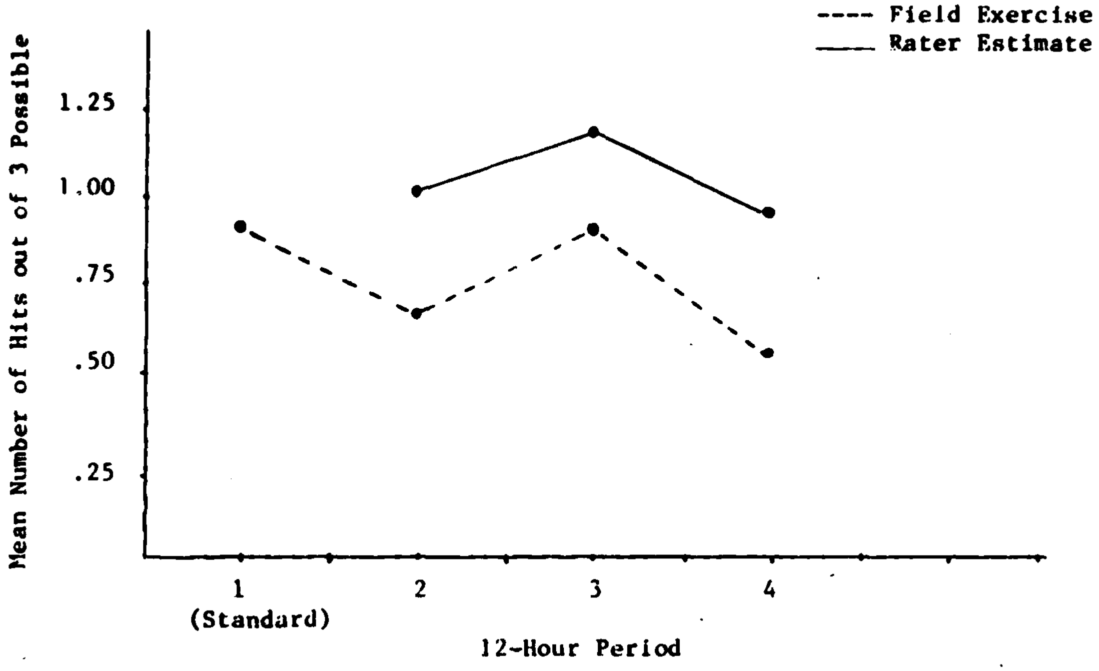

Figure 1. Mean hit scores with main gun for successive 12-hour periods with tanks and targets stationary **(Armor)**.

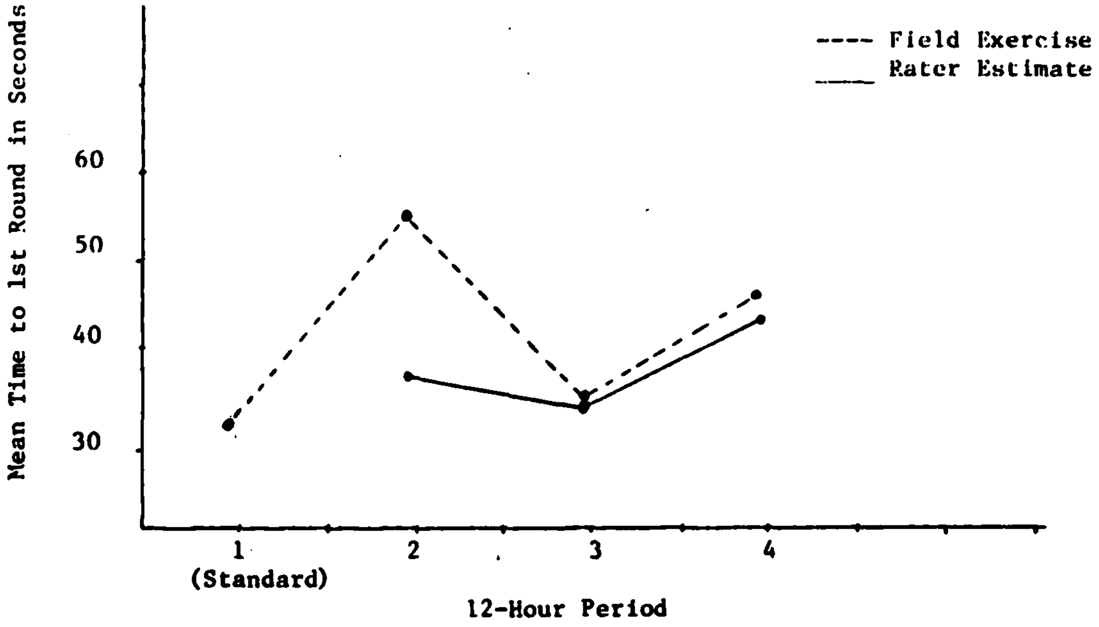

Figure 2. Mean time to fire first round with main gun for successive 12-hour periodswith tanks and targets stationary **(Armor)**.

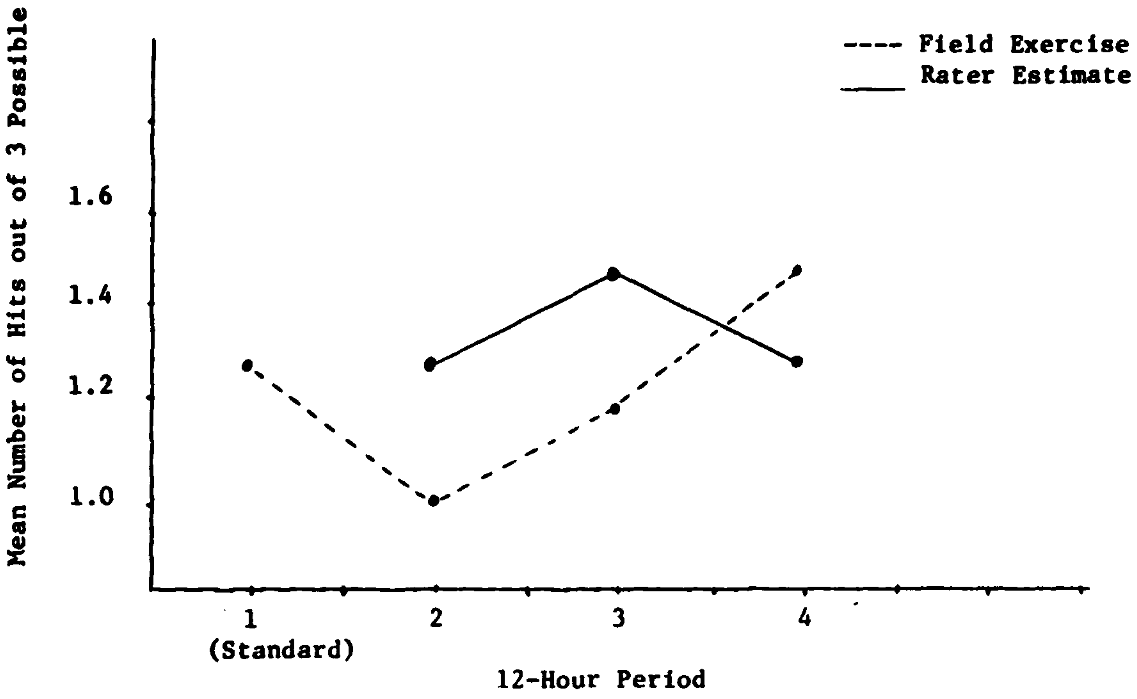

Figure 3. Mean number of hits per three rounds with main gun for successive 12-hour periods (tanks stationary, targets moving) **(Armor)**.

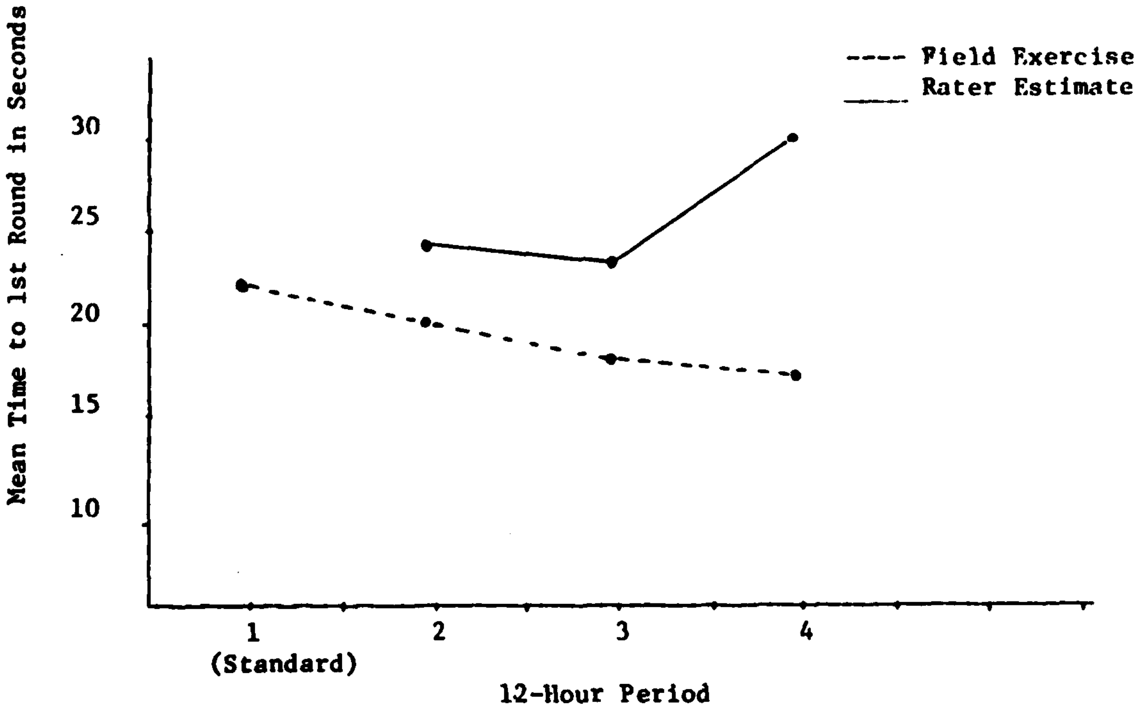

Figure 4. Mean time to fire first round with main gun for successive 12-hour periods (tanks stationary, targets moving)' **(Armor)**.

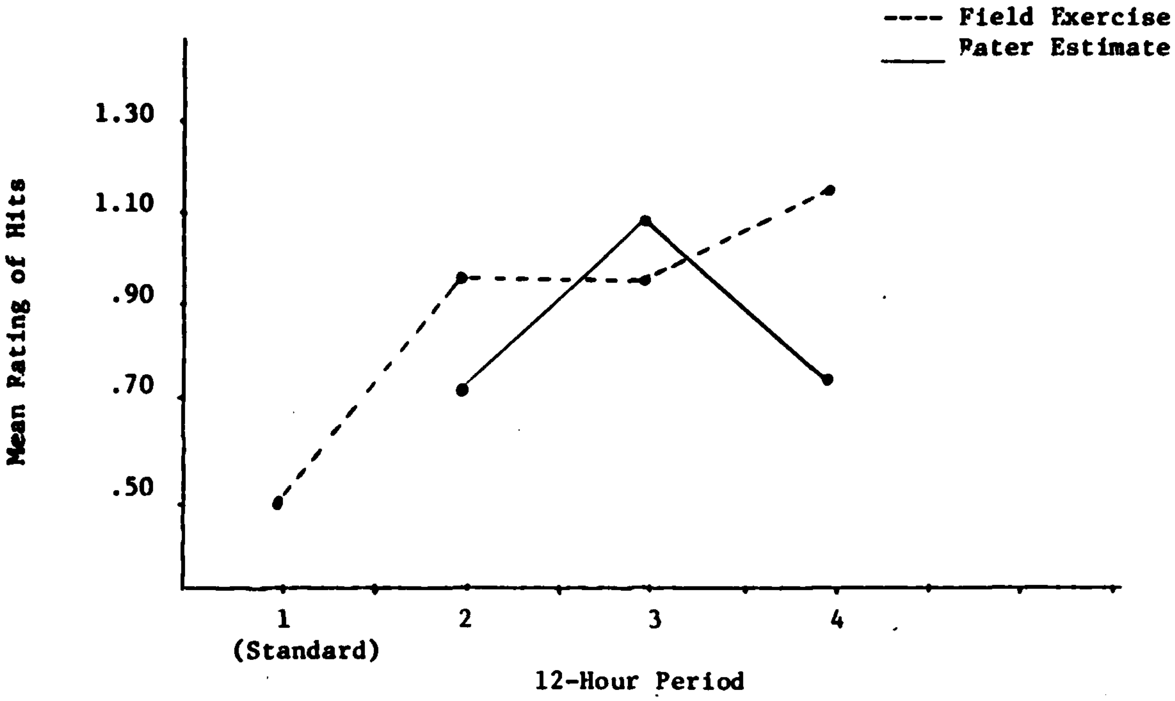

Figure 5. Mean ratings of hits per 50 rounds with caliber .50 machine gun for successive 12-hour periods (tanks stationary, targets moving) (Armor).

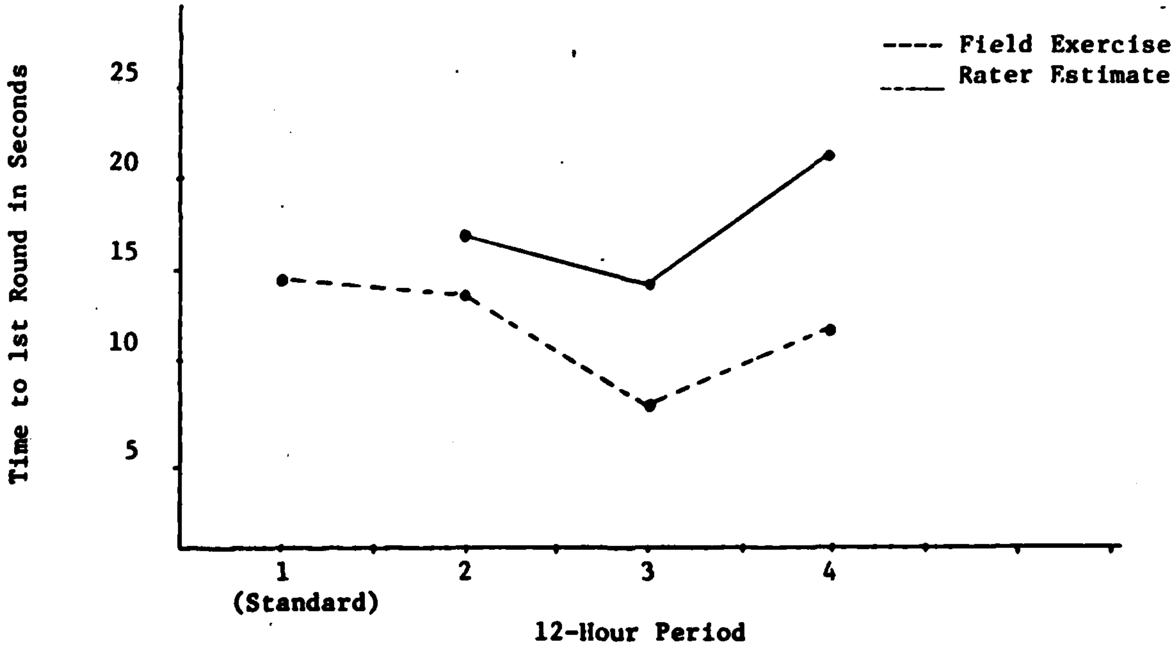

Figure 6. Mean time to fire caliber .50 machine gun for successive 12-hour periods(tanks stationary, targets moving) (Armor).

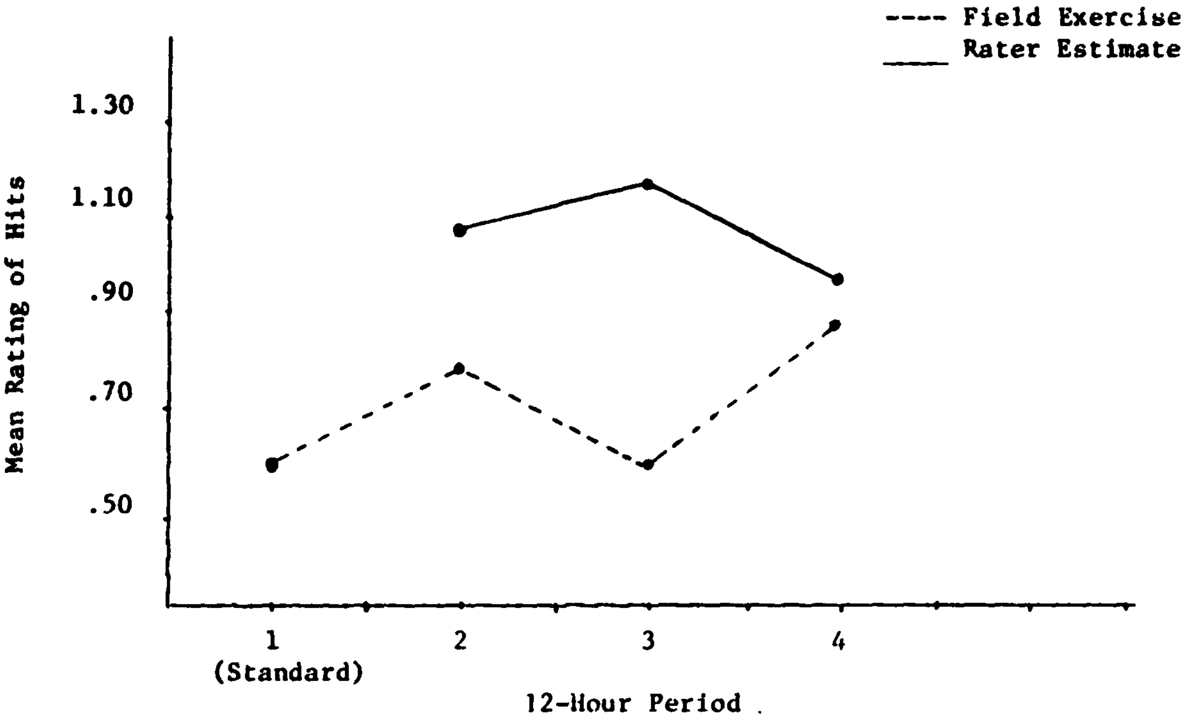

Figure 7. Mean ratings of hits per 20-round burst with coaxial machine gun for successive 12-hour periods (tanks moving, target stationary) (Armor).

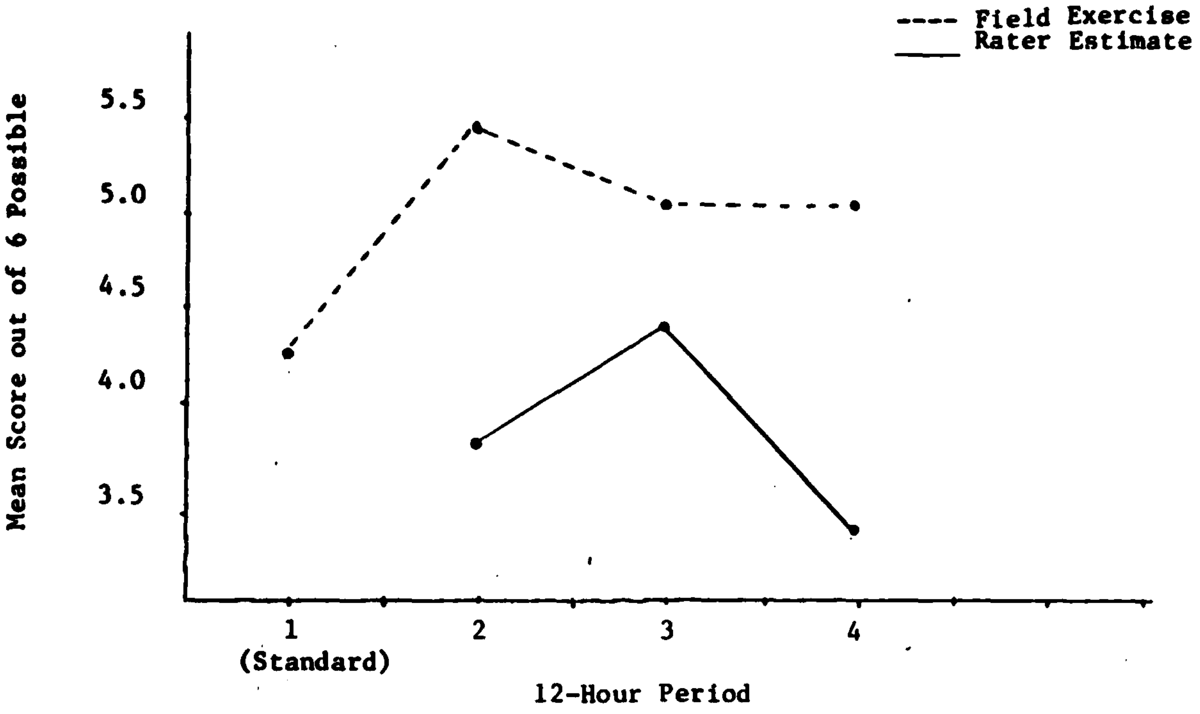

Figure 8. Mean log obstacle accuracy scores for successive 12-hour periods (Armor).

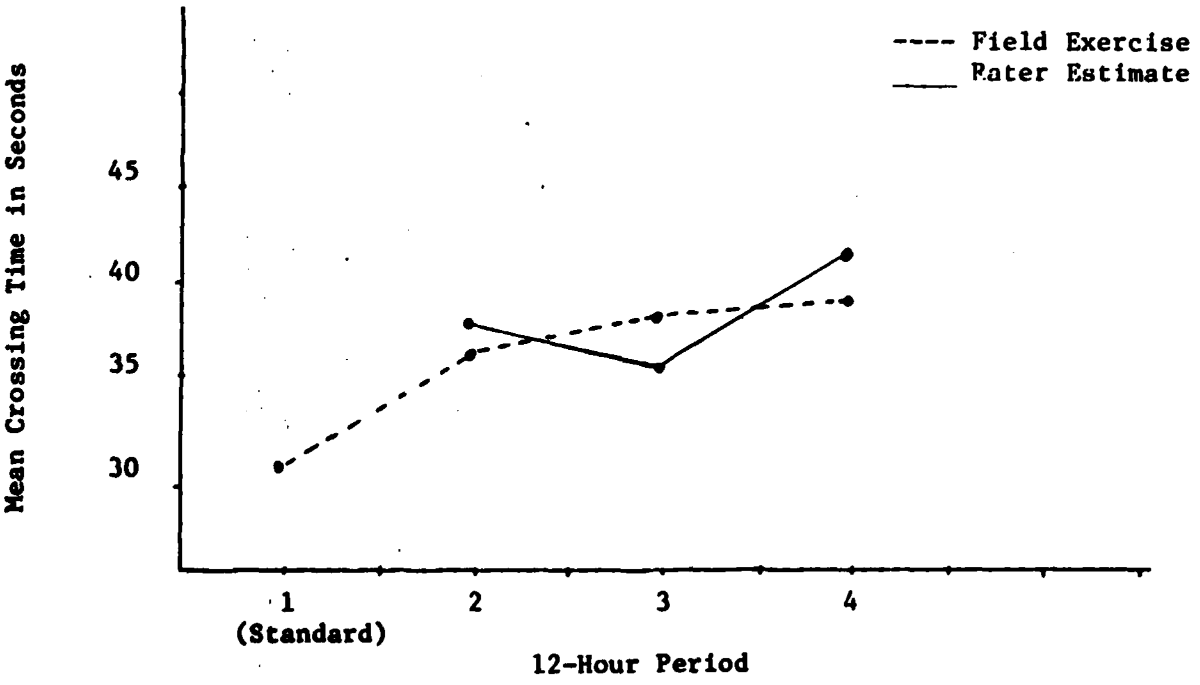

Figure 9. Mean log obstacle crossing times for successive 12-hour periods (Armor).

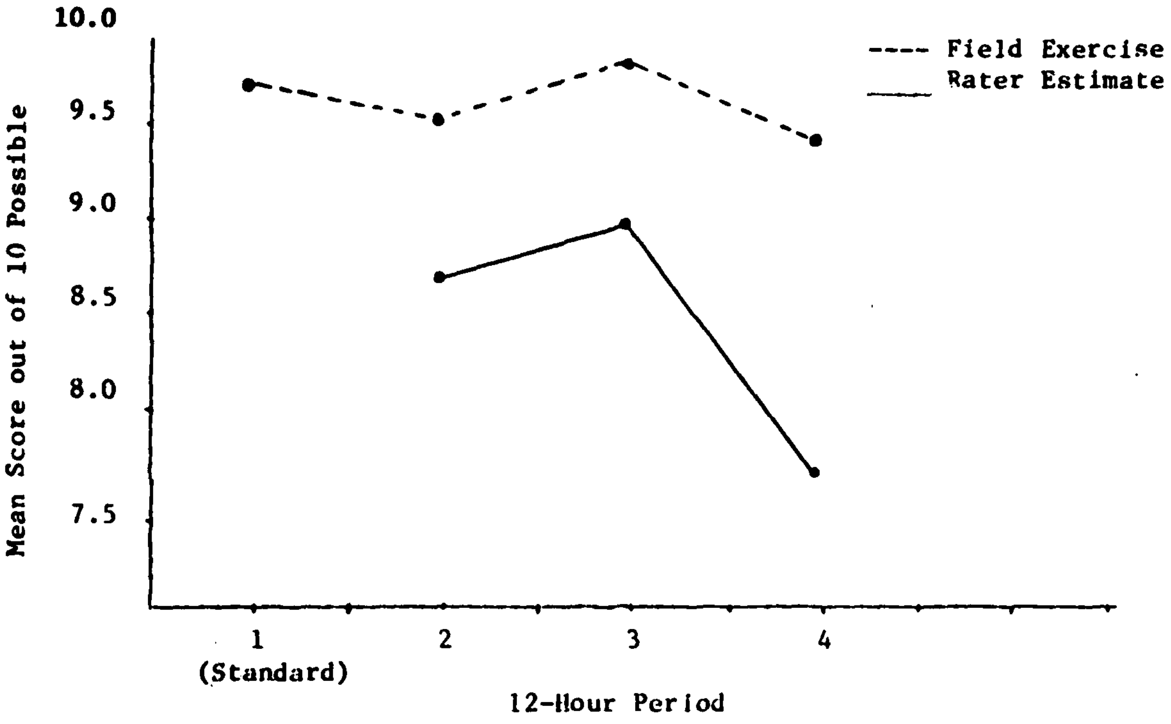

Figure 10. Mean slalom accuracy scores for successive 12-hour periods (Armor).

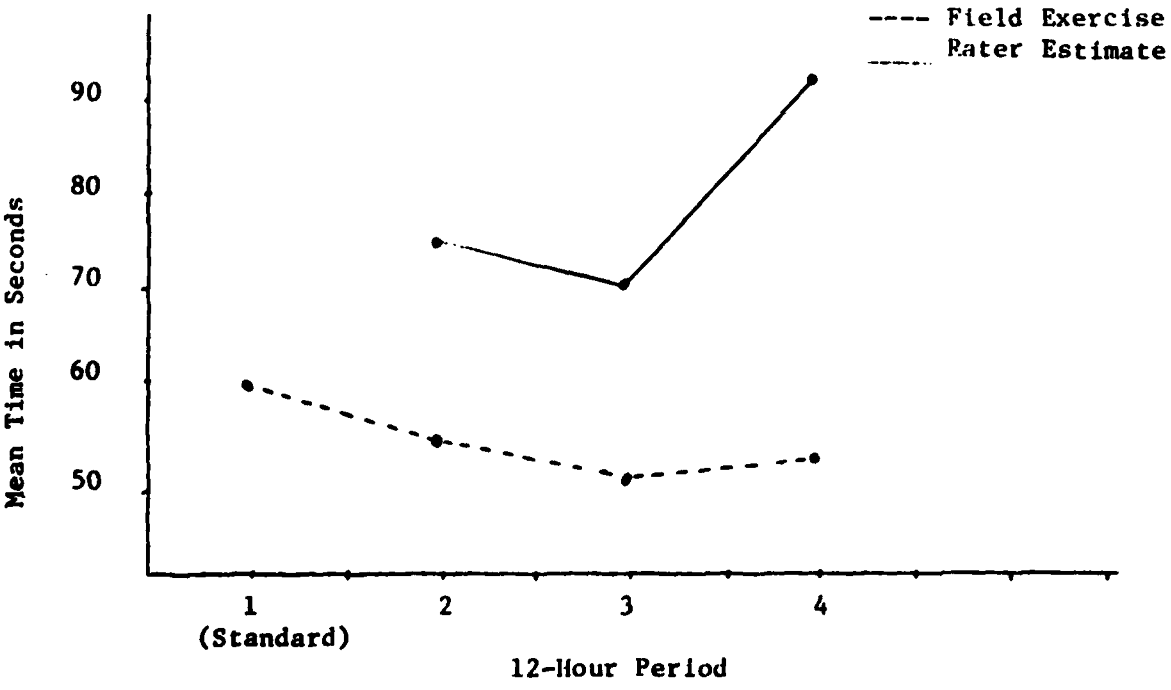

Figure 11. Mean slalom course times for successive 12-hour periods (Armor).

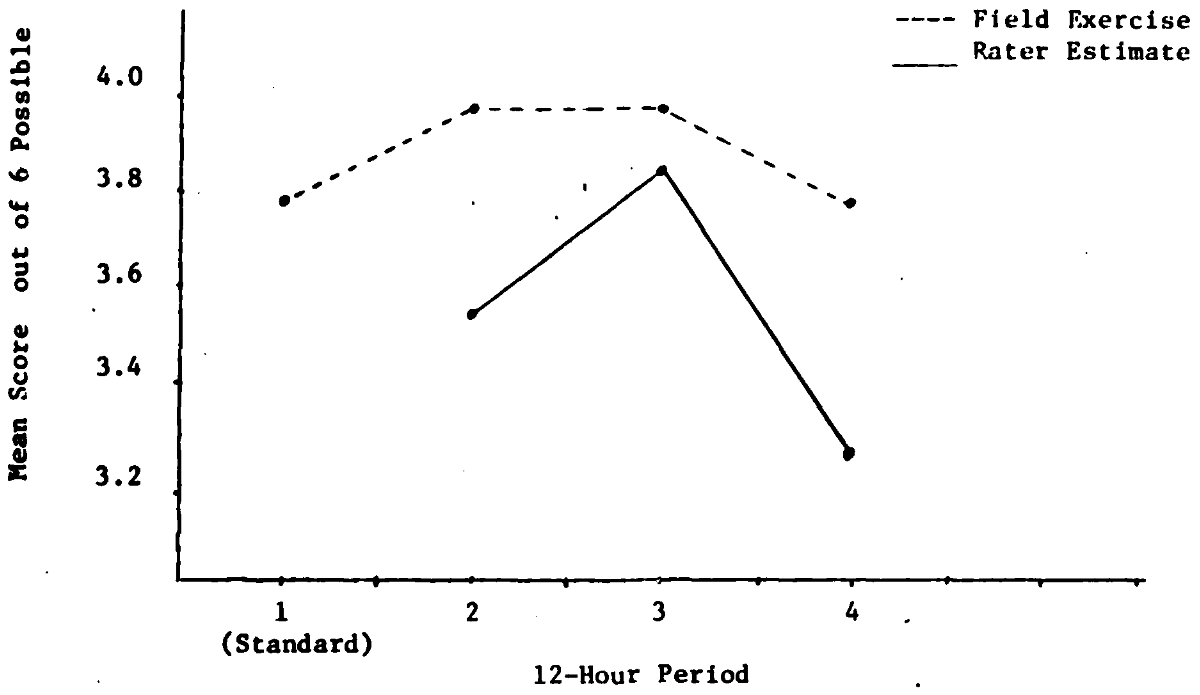

Figure 12. Mean ditch crossing accuracy scores for successive 12-hour periods (Armor).

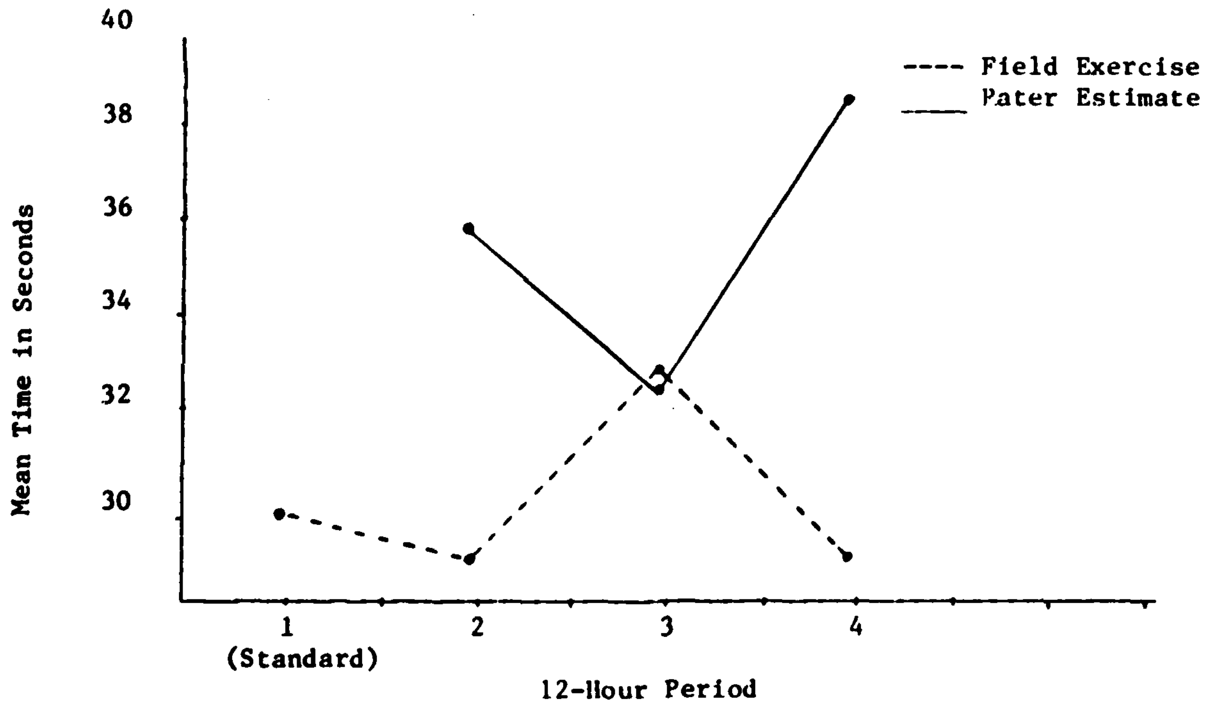

Figure 13. Mean ditch crossing times for successive 12-hour periods (Armor).

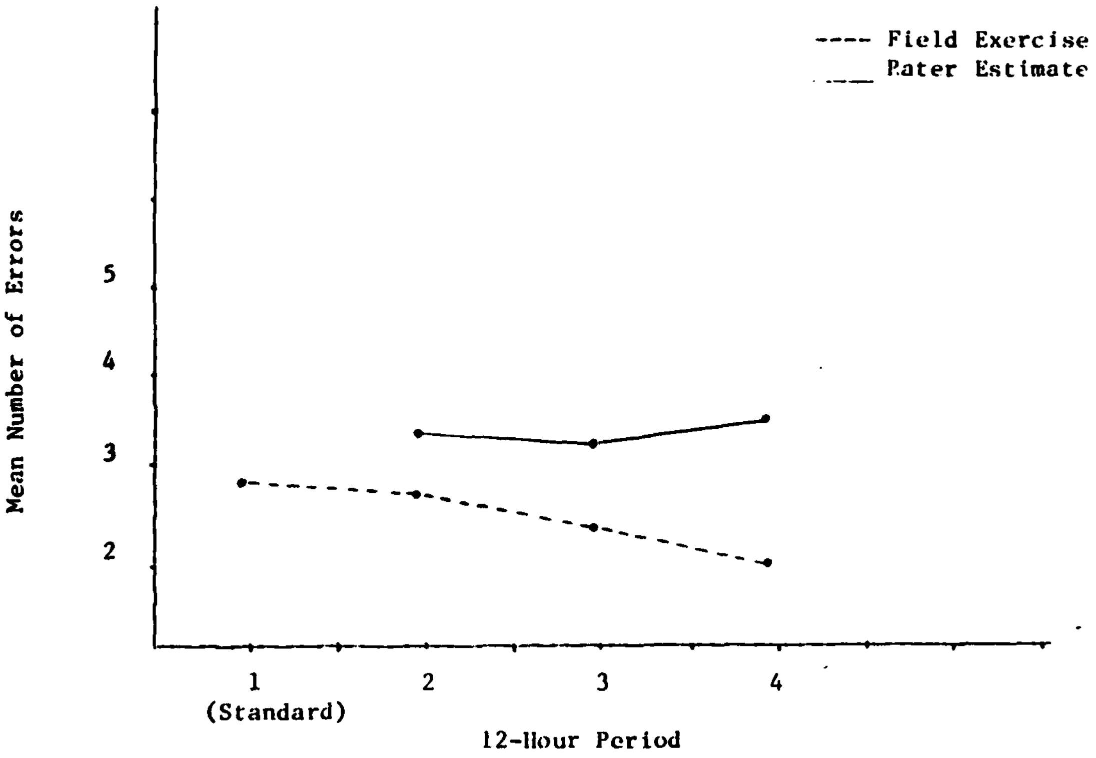

Figure 14. Mean minefield crossing accuracy scores for successive 12-hour periods (Armor).

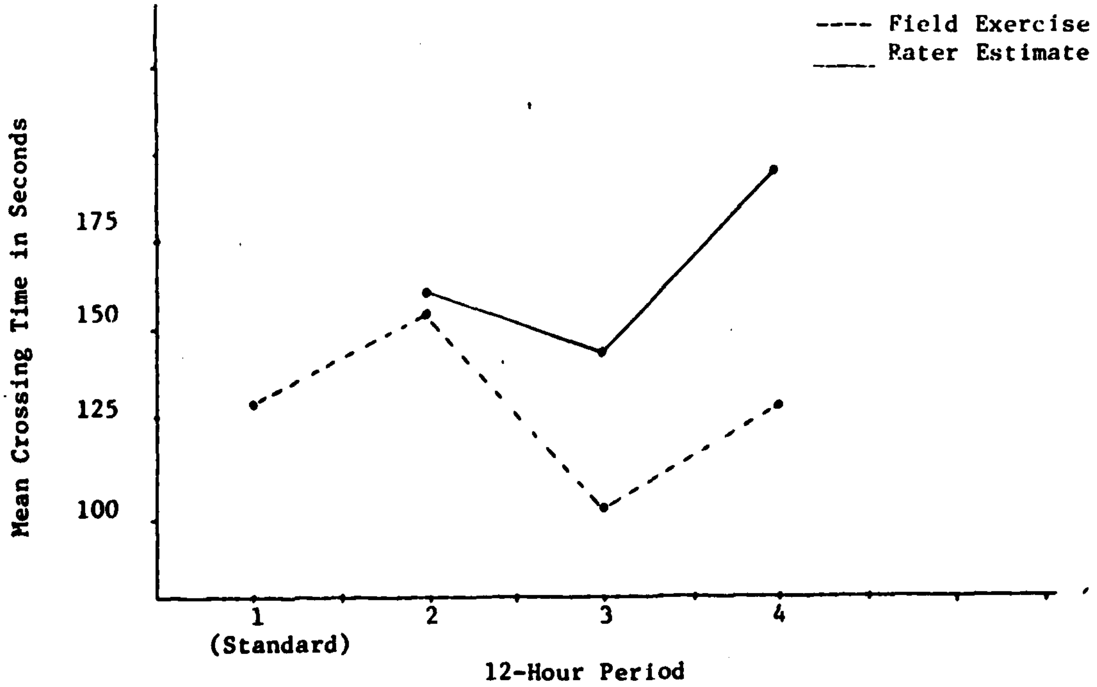

Figure 15. Mean minefield crossing times for successive 12-hour periods (Armor).

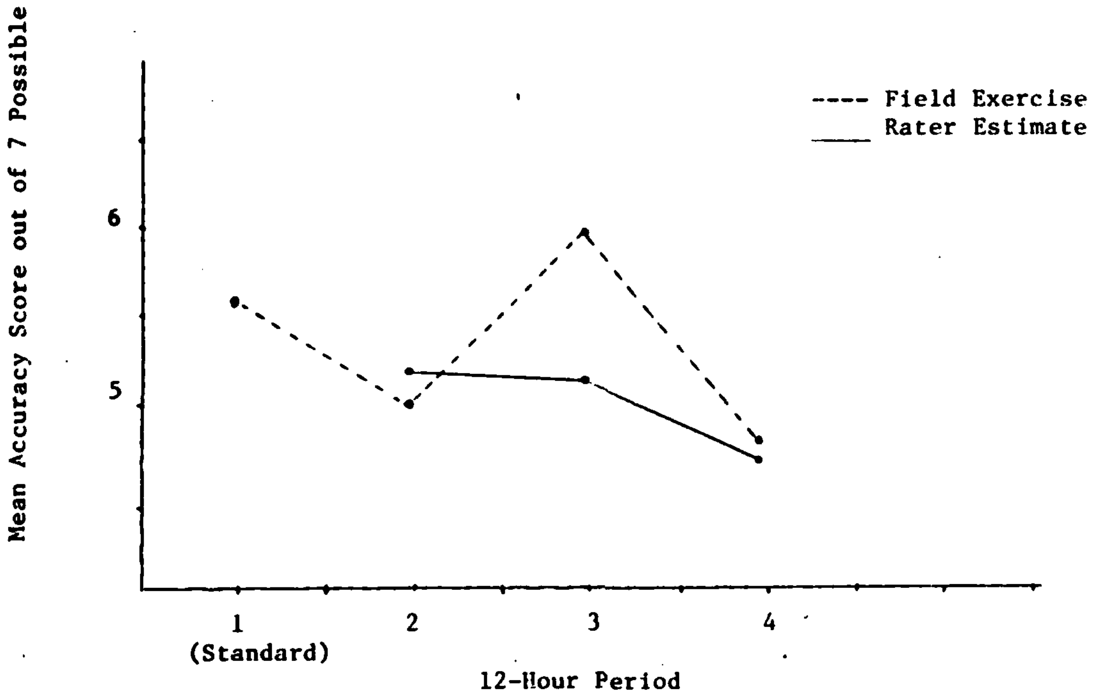

Figure 16. Mean number of maintenance tasks correctly performed for successive 12-hour periods (Armor).

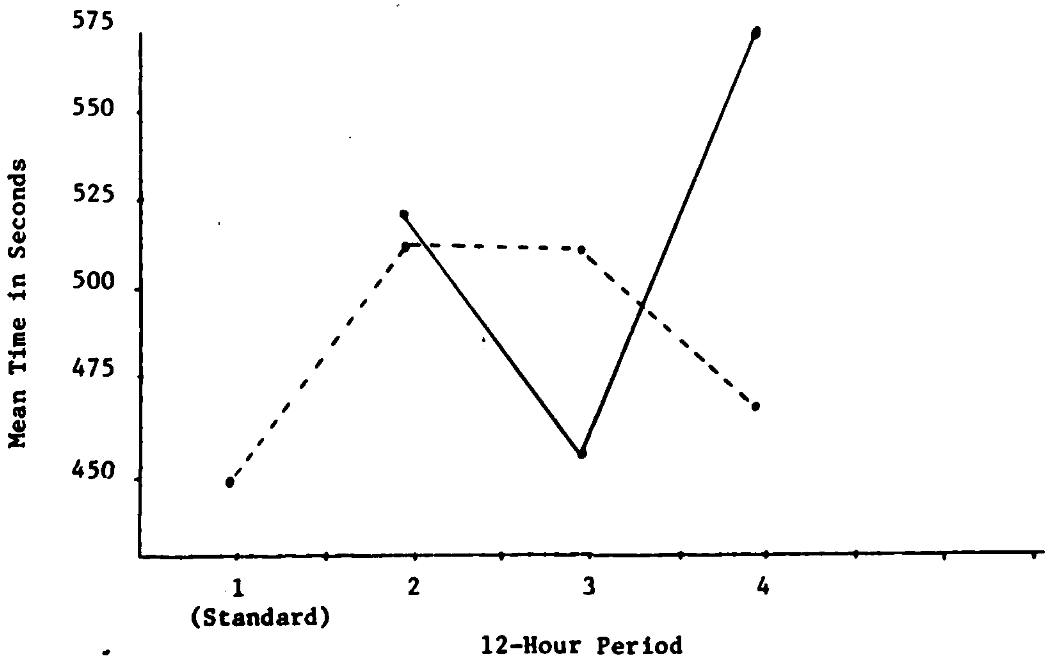

Figure 17. Mean times for performance of maintenance tasks for successive 12-hour periods (Armor).

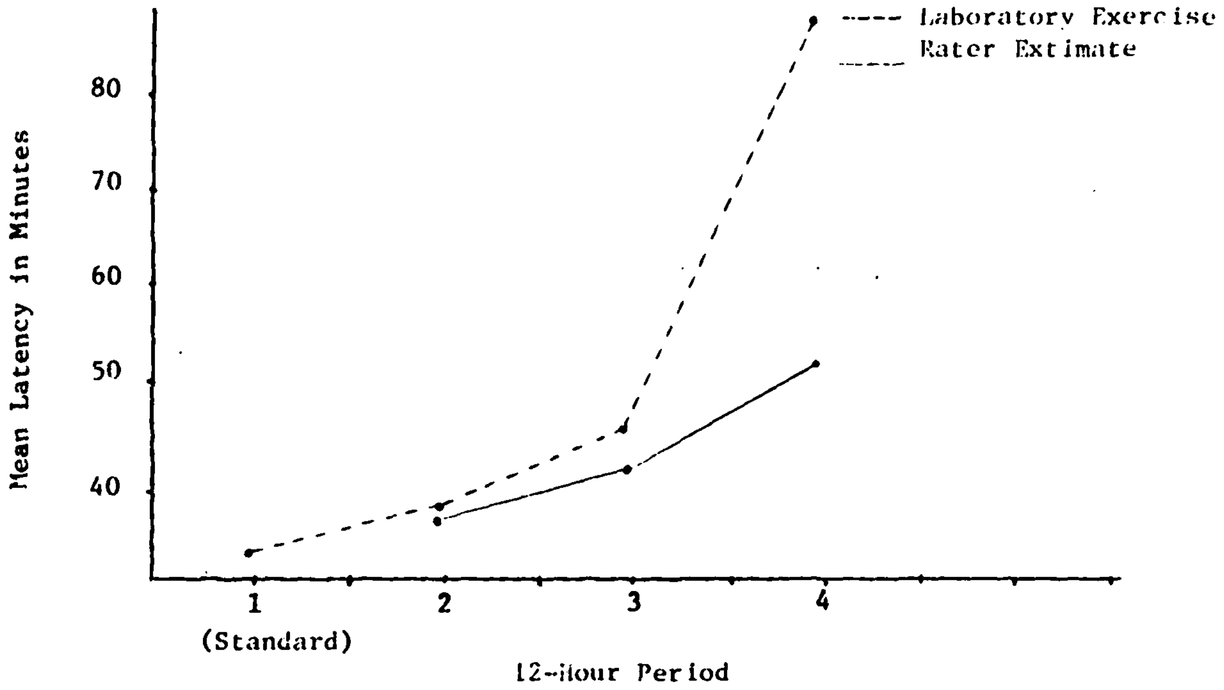

Figure 18. Mean latency to completion of preplanned tasks for successive 12-hour periods (FDC).

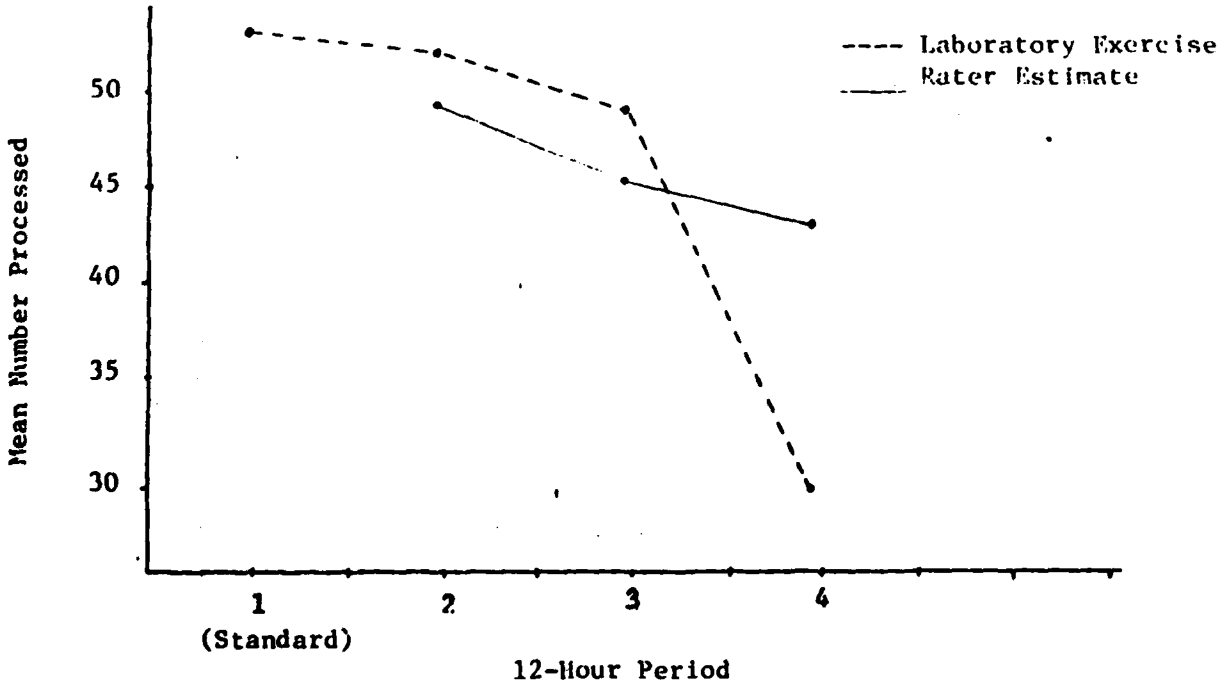

Figure 19. Mean number of preplanned missions processed for successive 12-hour periods (FDC).

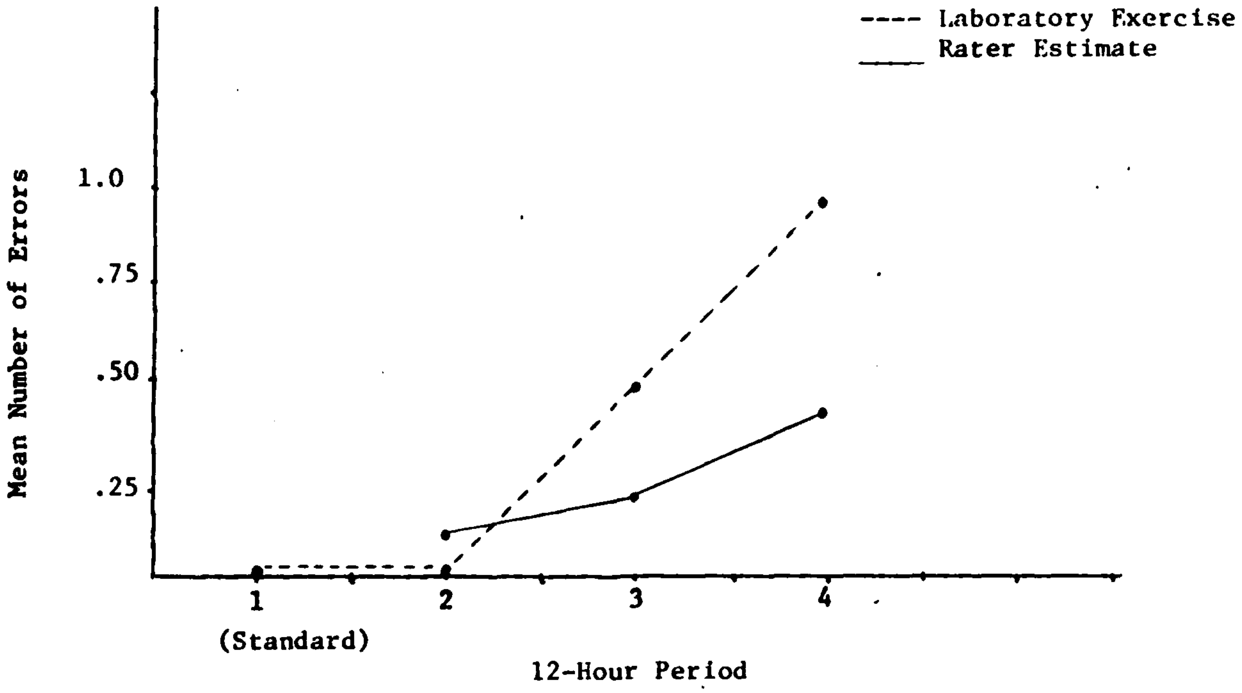

Figure 20. Mean number of errors greater than 990 mils for preplanned tar'gets for successive 12-hour periods (FDC).

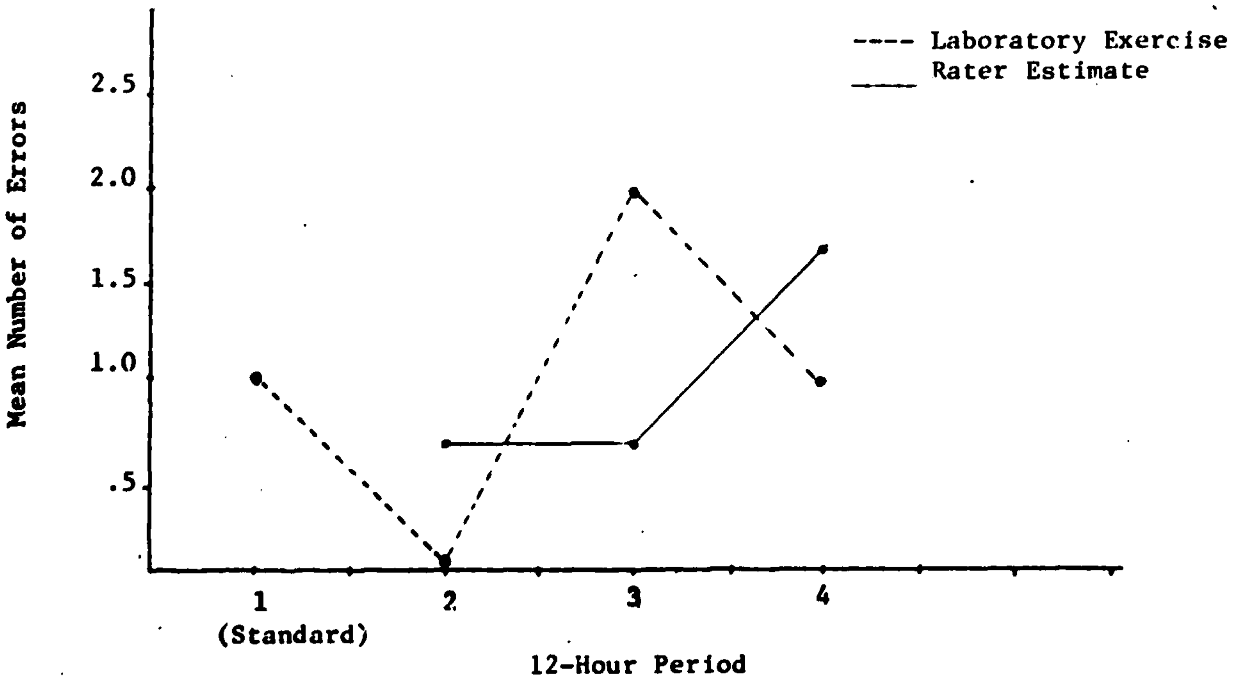

Figure 21. Mean number of errors 90-990 mils for preplanned targets for successive 12-hour periods (FDC).

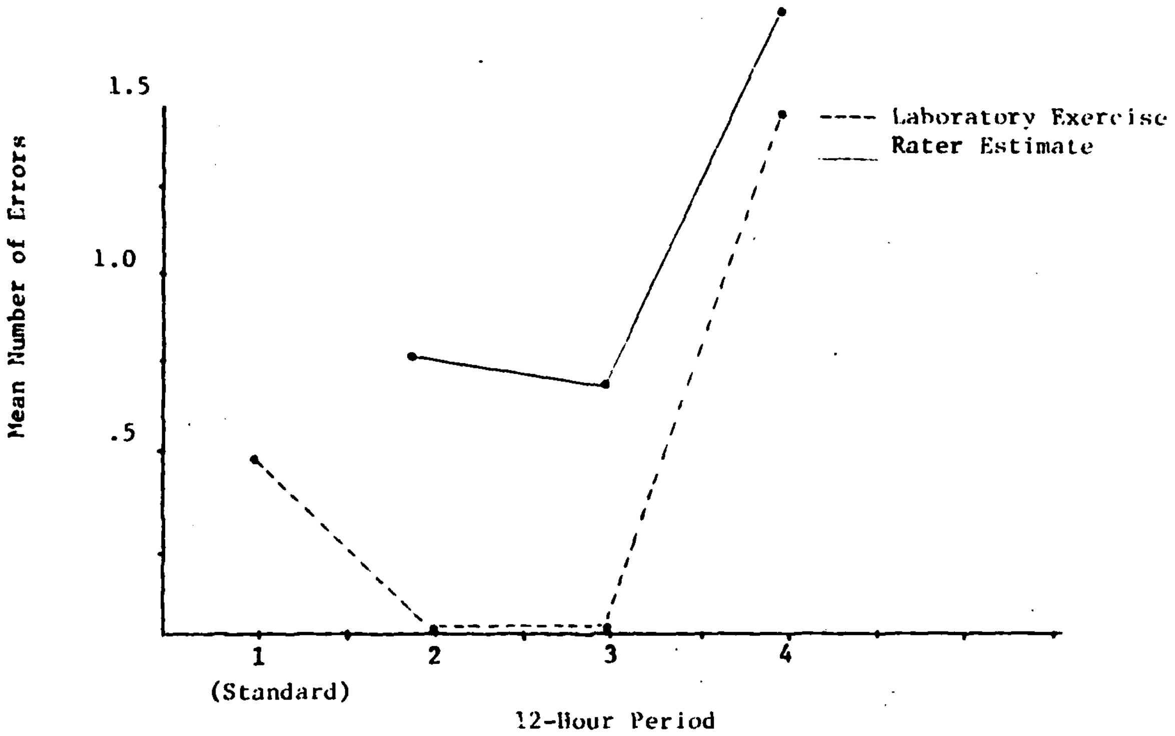

Figure 22. Mean number of errors 30 to 89 mils for preplanned targets for successive 12-hour periods (FDC).

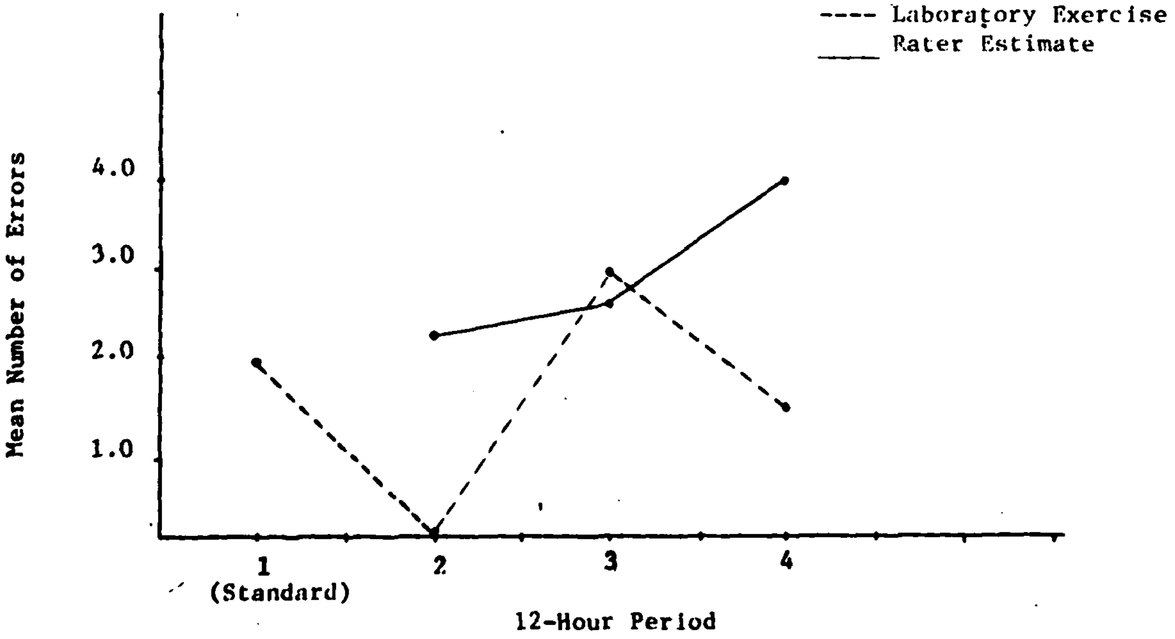

Figure 23. Mean number of errors 15-29 mils for preplanned targets for successive 12-hour periods (FDC).

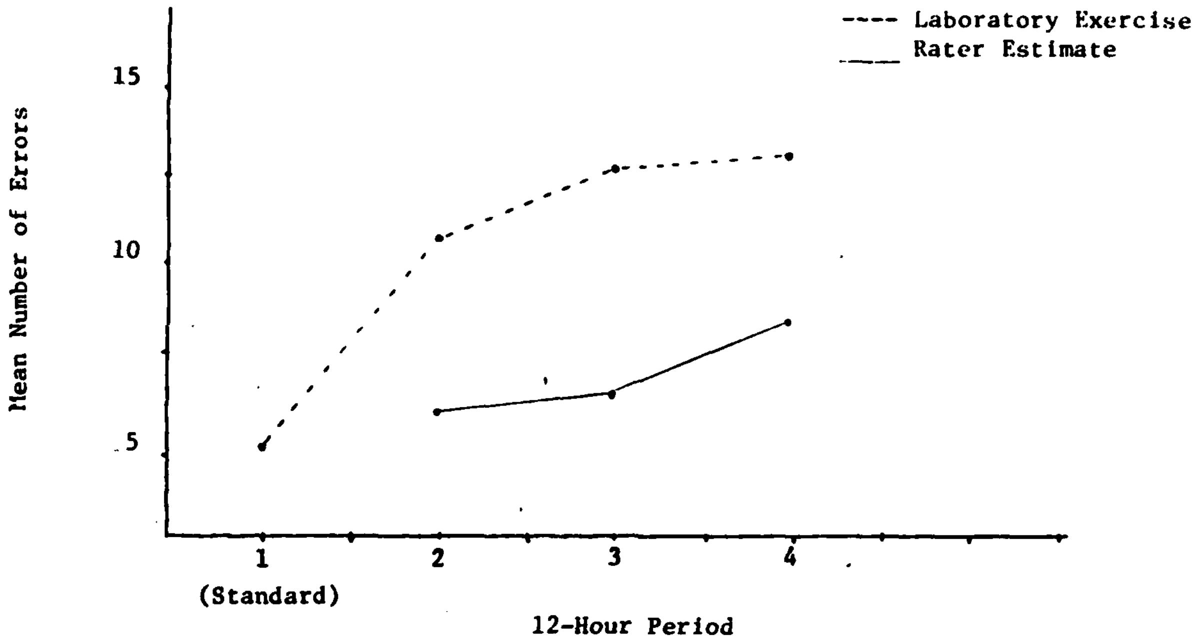

Figure 24. Mean number of errors 7-14 mils for preplanned targets for successive 12-hour periods (FDC).

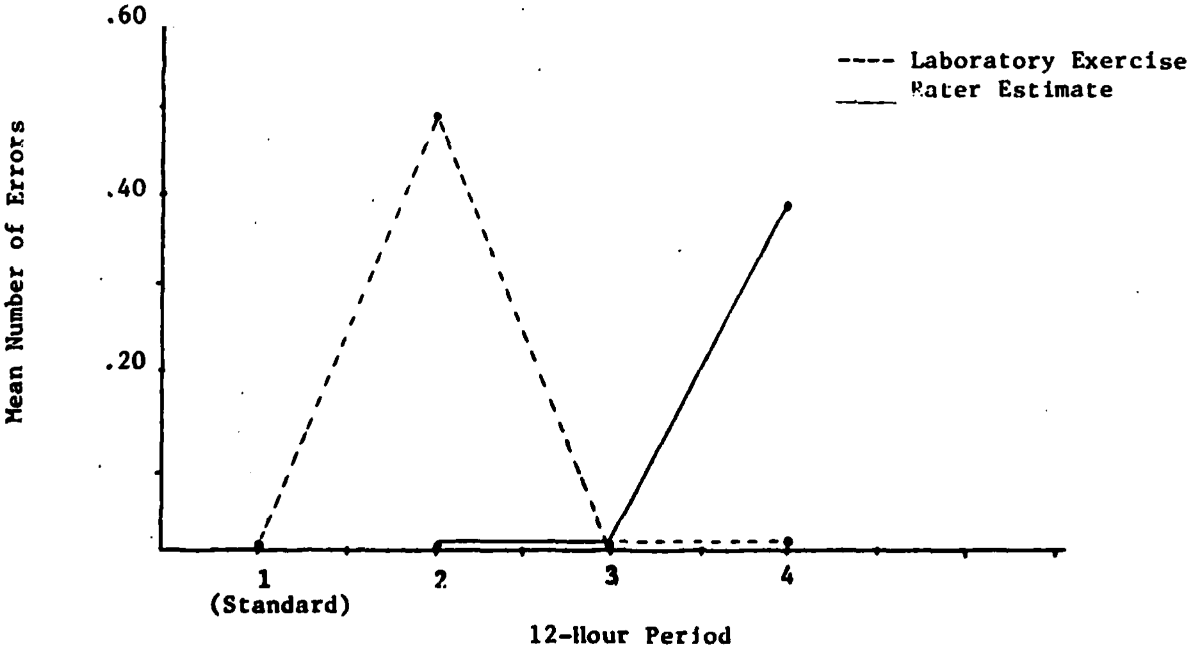

Figure 25. Mean number of errors greater than 990 mils for unplanned missions or subsequent adjustments for successive 12-hour periods (FDC).

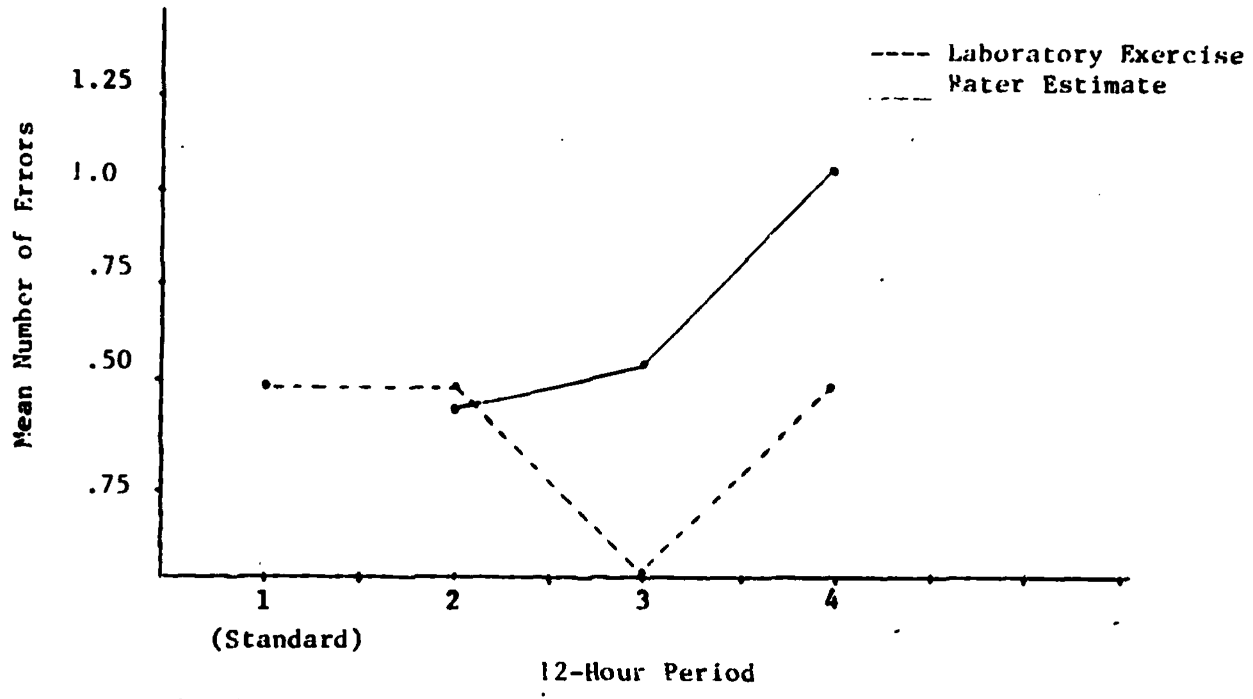

Figure 26. Mean number of errors 90-990 mils for unplanned missions or subsequent adjustments for successive 12-hour periods (FDC).

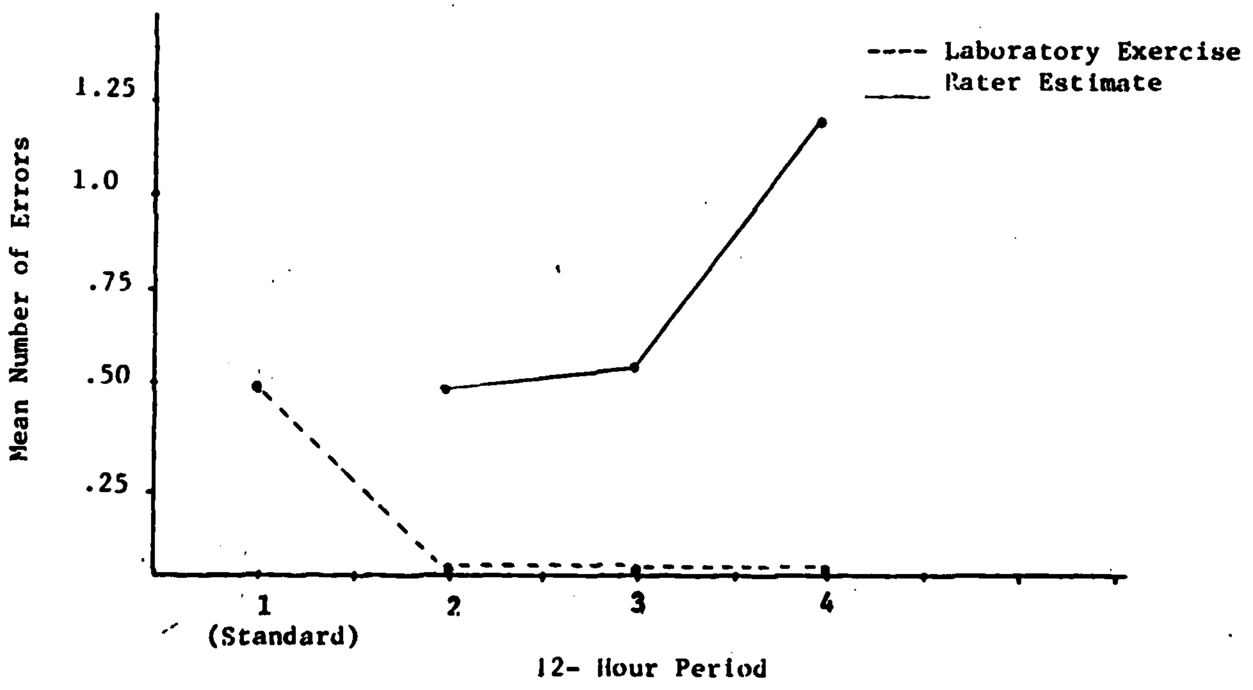

Figure 27. Mean number of errors 30-89 mils for unplanned missions or subsequent adjustments for successive 12-hour periods (FDC).

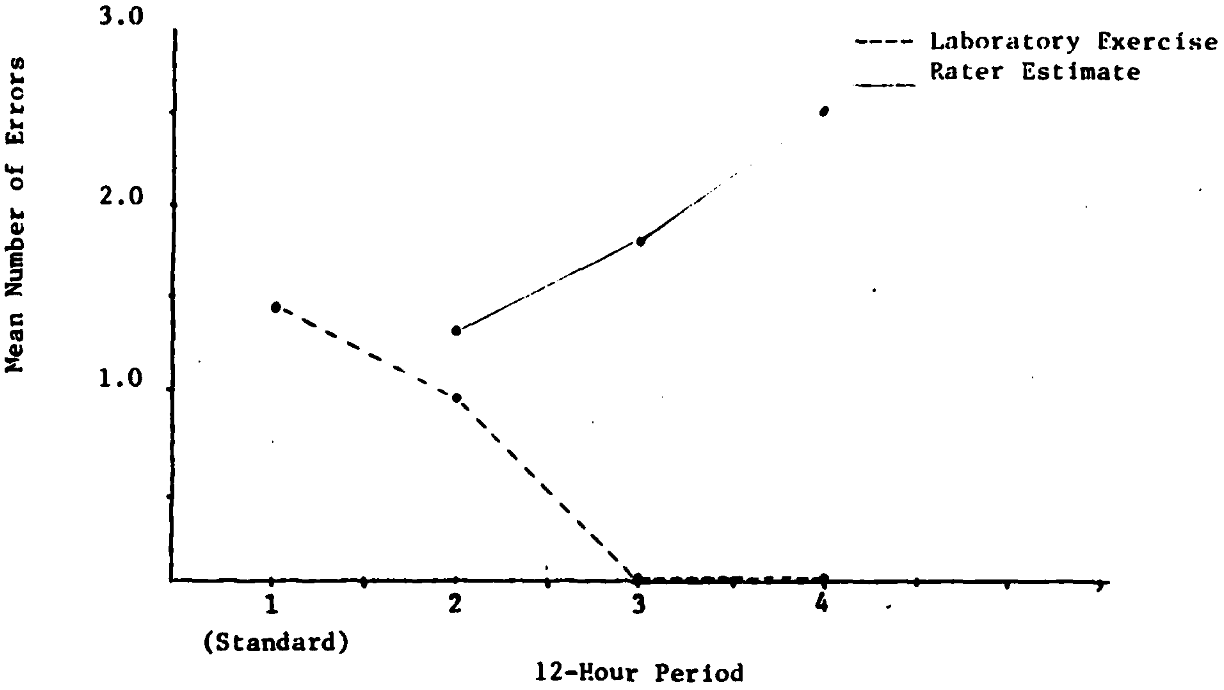

Figure 28. Mean number of errors 15-29 mils for unplanned missions or subsequent adjustments for successive 12-hour periods (FDC).

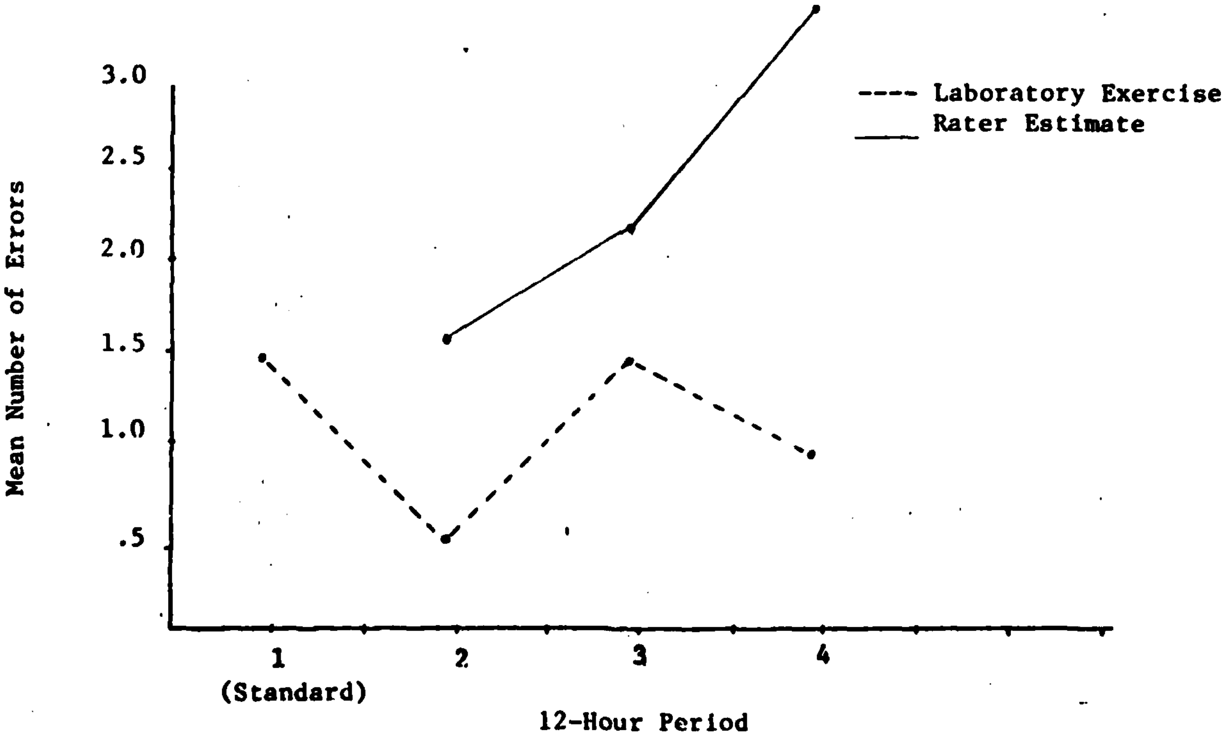

Figure 29. Mean number of errors 7-14 mils for unplanned missions or subsequent adjustments for successive 12-hour periods (FDC).

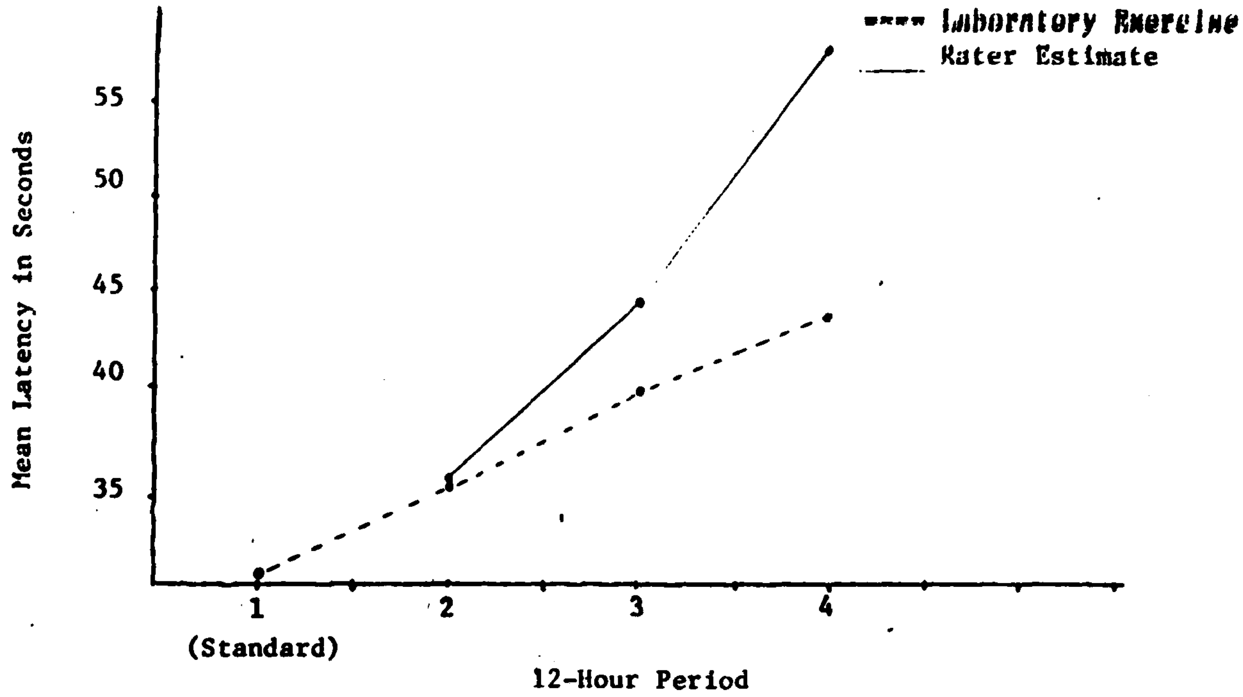

Figure 30. Mean computation latencies for unplanned missions for successive 12-hour periods (FDC).

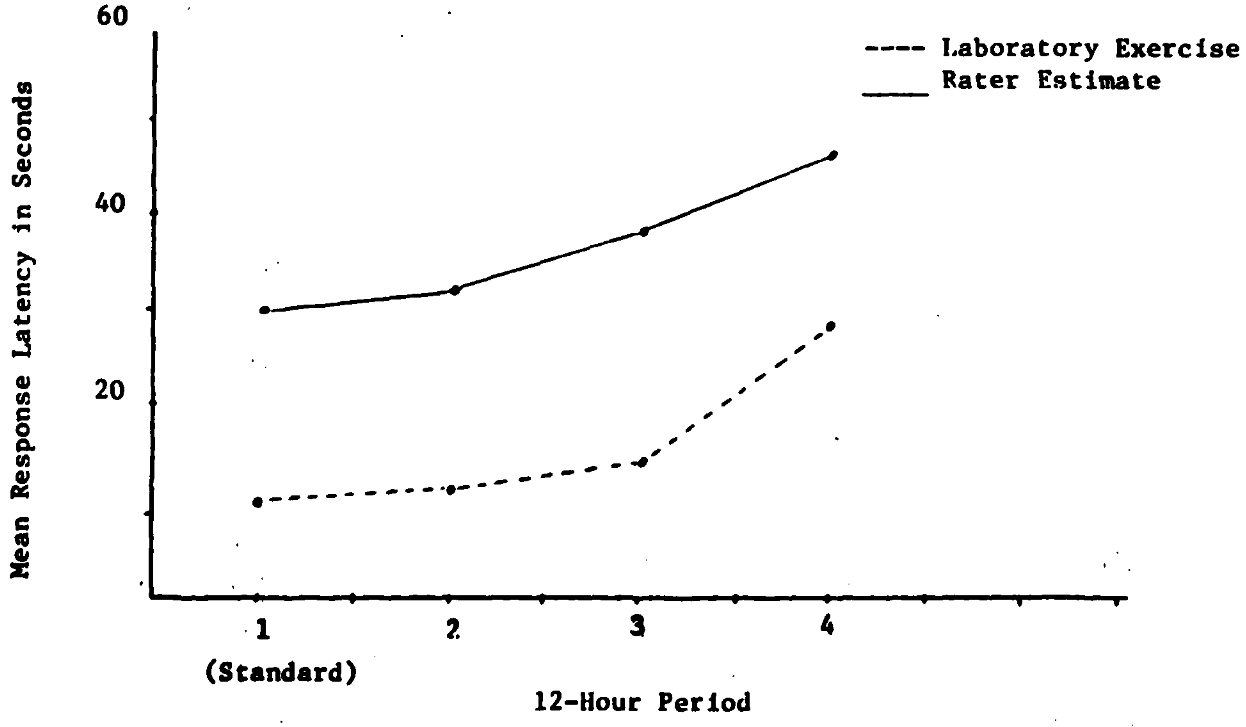

Figure 31. Mean time to respond a once call for on-call mission was made, for successive 12-hour periods (FDC).

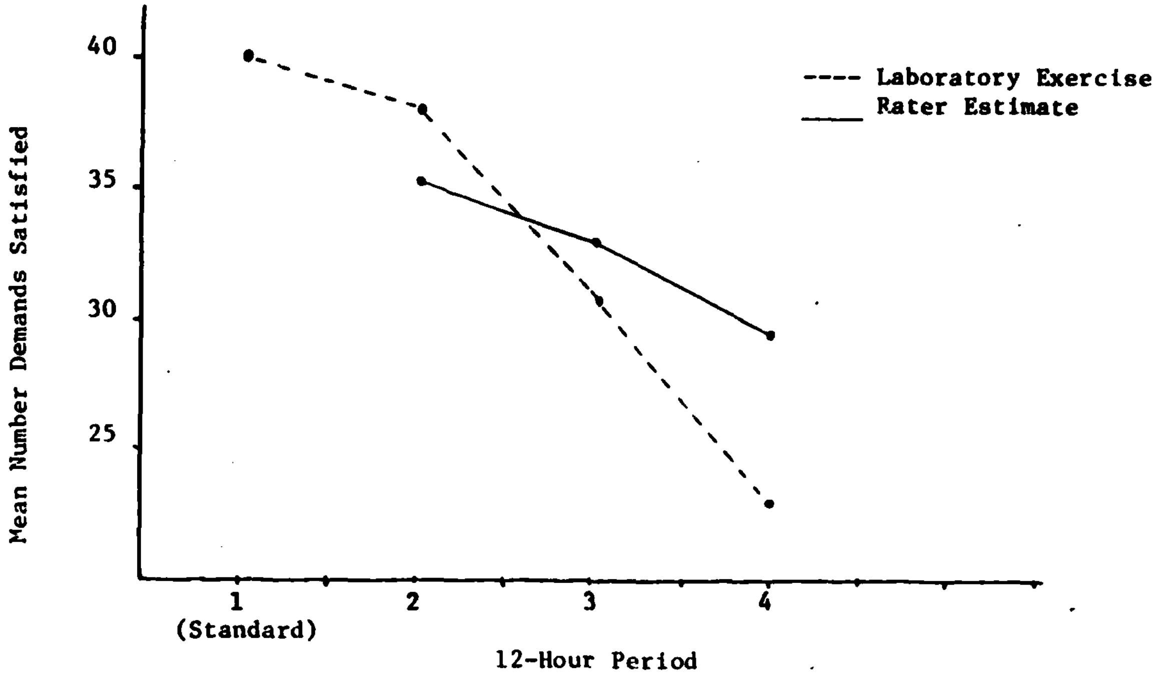

Figure 32. Mean number of prioritizing demands satisfied (maximum possible 44) for successive 12-hour periods (includes completion of precomputations of firing data for prioritized targets and designation of the target as priority) (FDC).

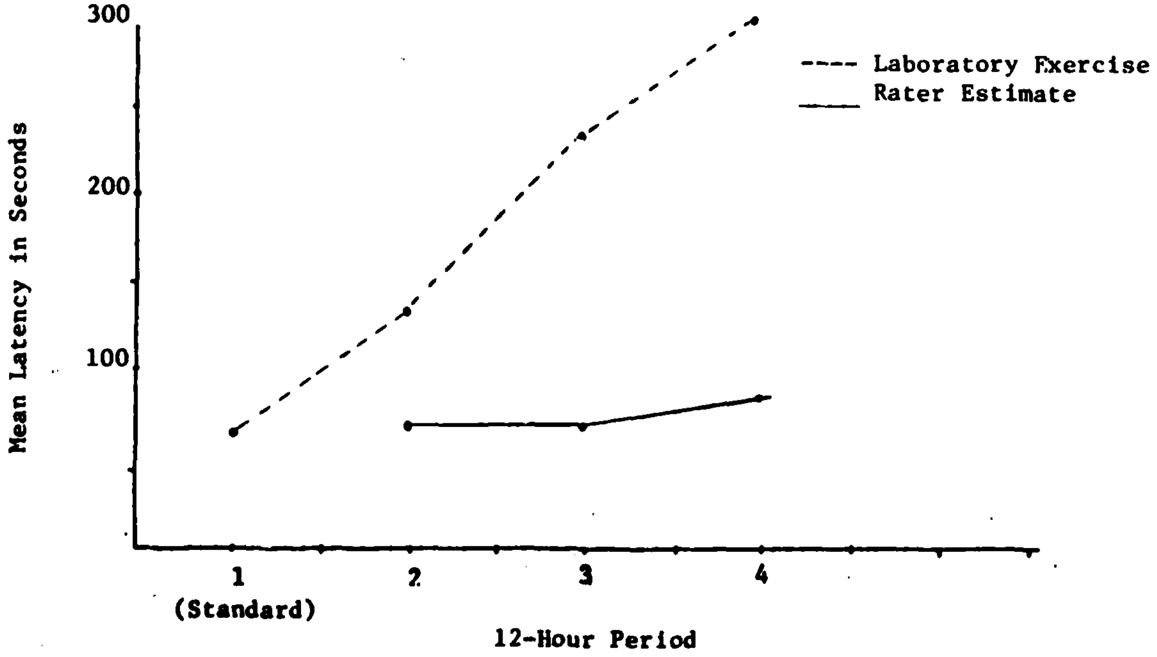

Figure 33. Mean prioritizing latency for successive 12-hour periods (FDC).

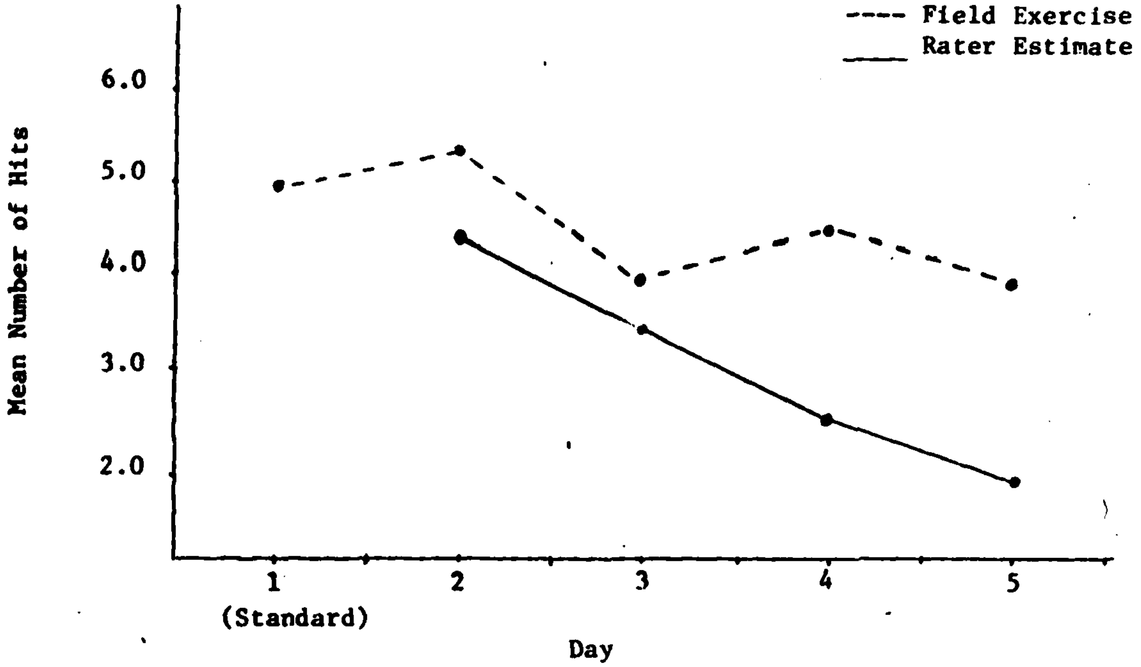

Figure 34. Mean number of hits out of nine rounds fired in vigilance shooting task for successive days (Early Call I).

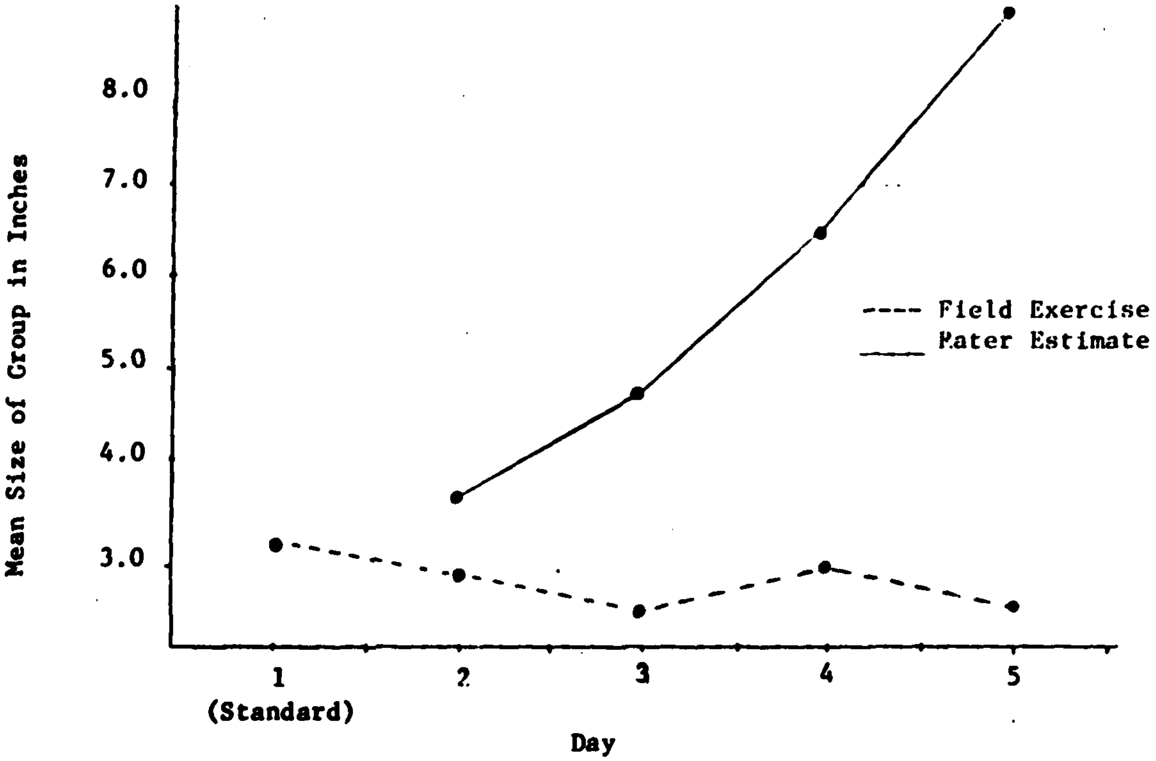

Figure 35. Mean size of best group of five rounds for successive days (Early Call I).

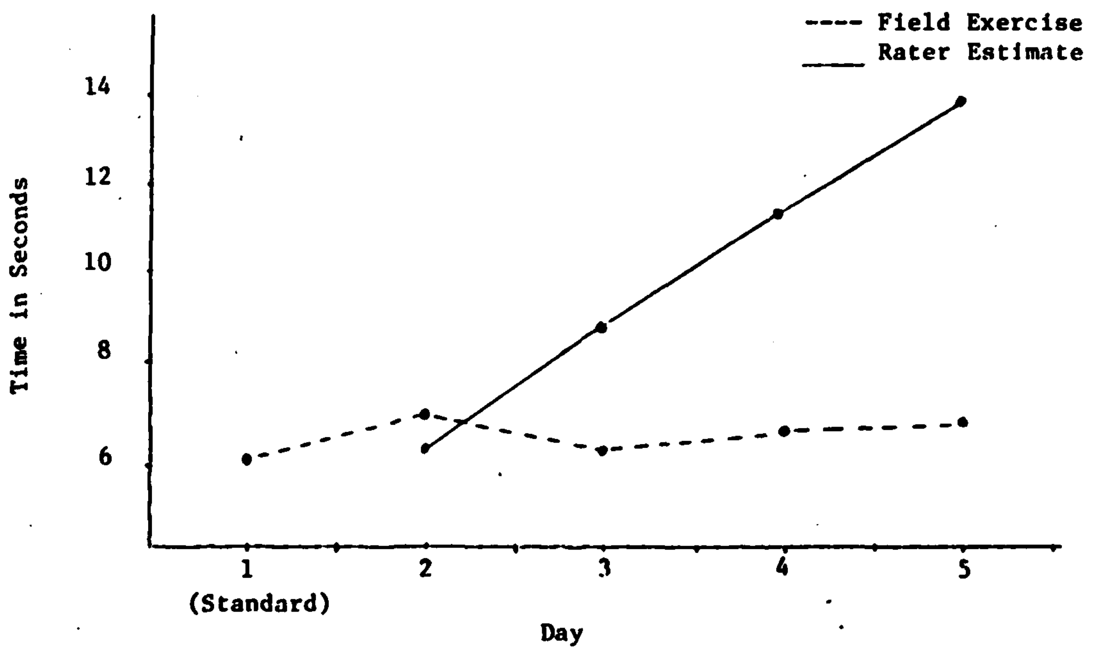

Figure 36. Time to load the rifle from the standing position for successive days in weapon-handling tests (Early Call I).

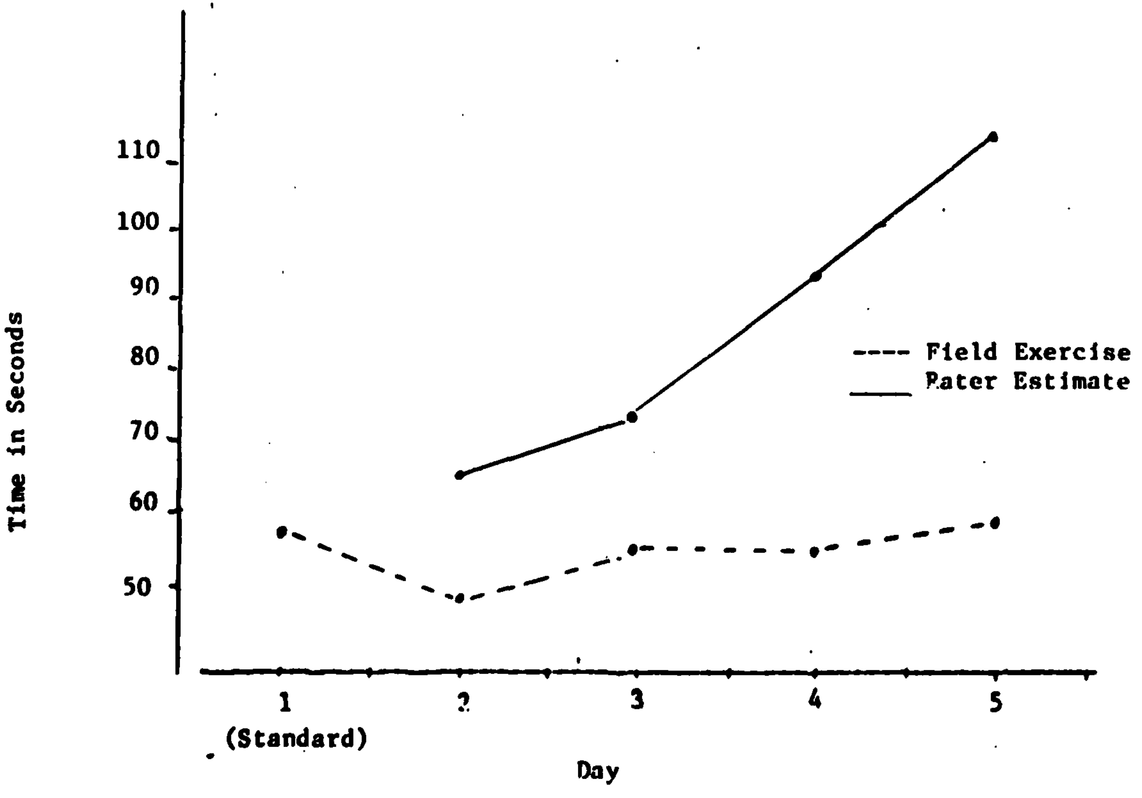

Figure 37. Time to assemble the rifle for successive days in the weapon-handling tests (Early Call I).

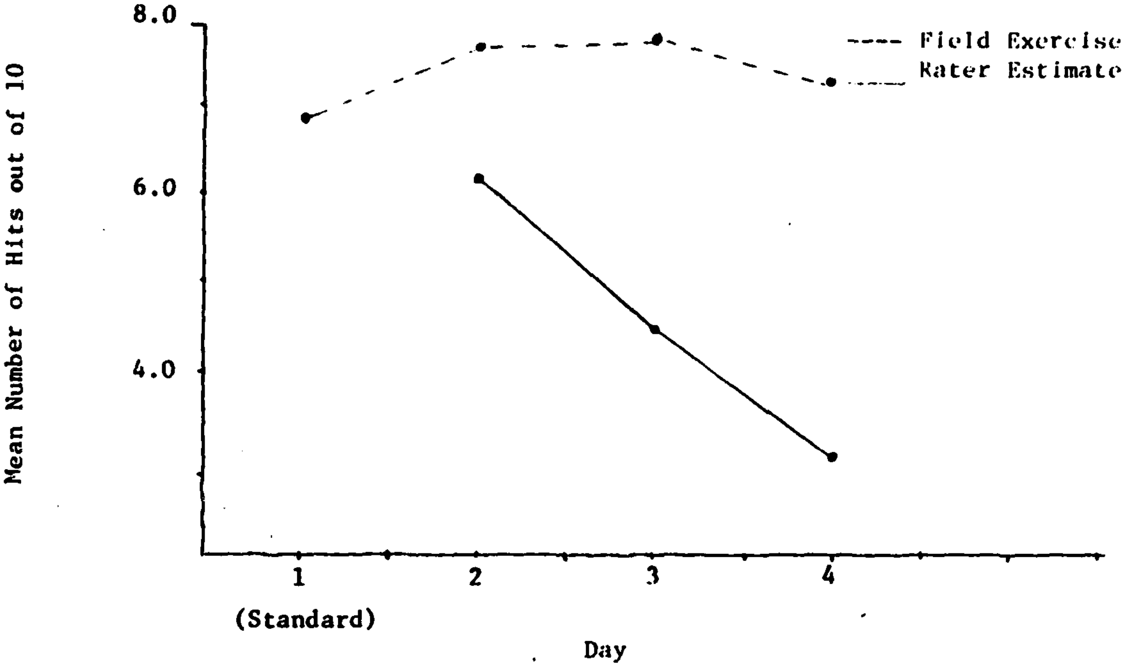

Figure 38. Mean number of hits in moving target test for successive days **(Early Call II).**

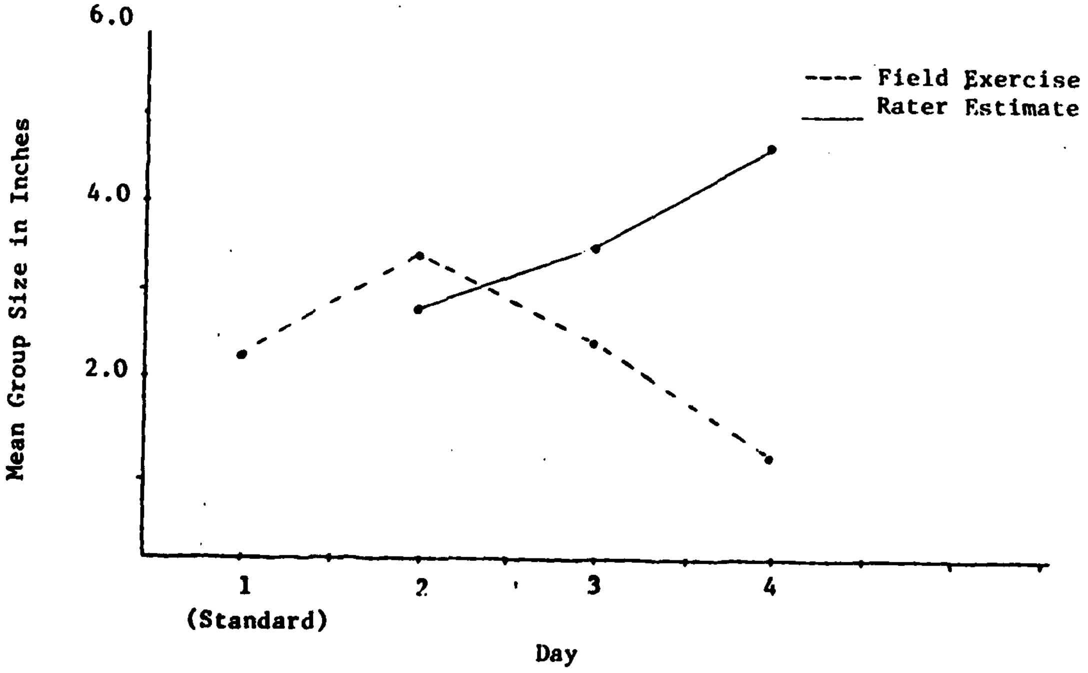

Figure 39. Mean group size, measured to the nearest quarter inch, for successive days **(Early Call II).**

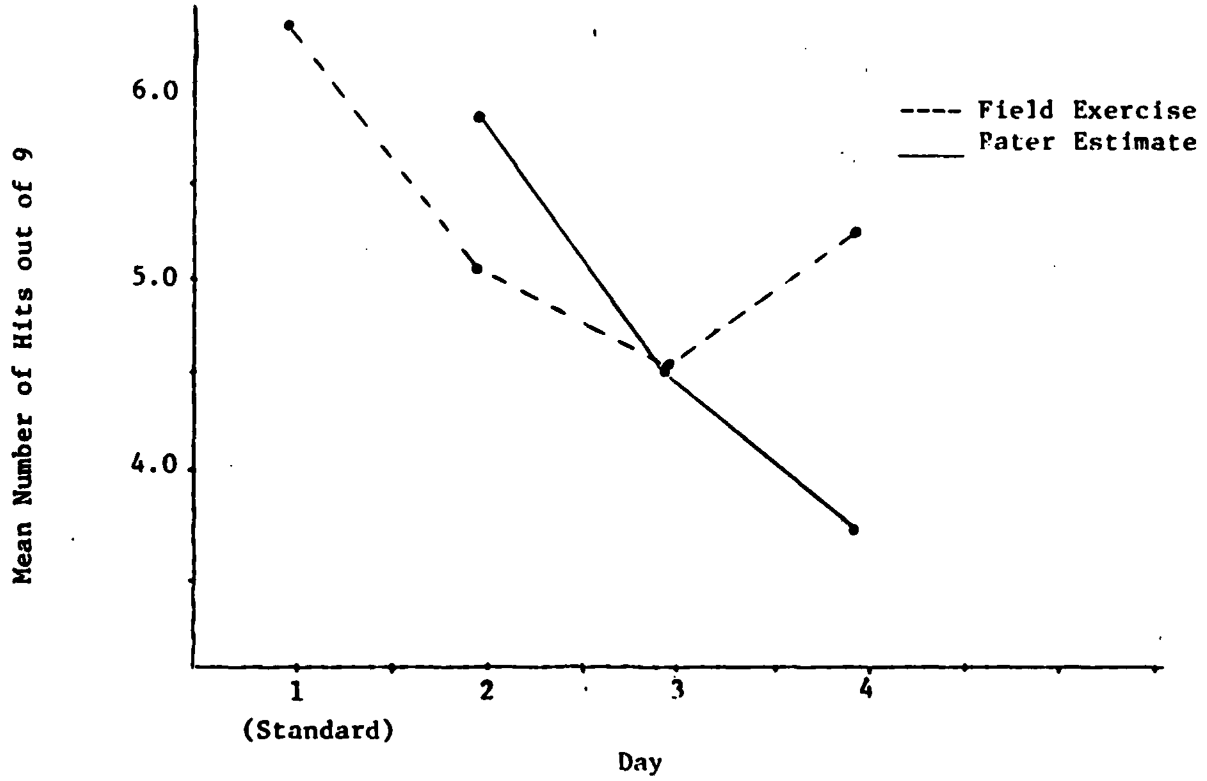

Figure 40. Mean number of hits scored in vigilance shooting test on successive days (Early Call II).