Research Note 84-53

THE VALIDITY OF PERFORMANCE SCORES ADJUSTED FOR ENVIRONMENTAL CONDITIONS

Laurel W. Oliver and Cecil D. Johnson

Submitted by

T. Owen Jacobs, Chief
LEADERSHIP AND MANAGEMENT TECHNICAL AREA

and

Joyce L. Shields, Director
MANPOWER AND PERSONNEL RESEARCH LABORATORY

U. S. Army

Research Institute for the Behavioral and Social Sciences

February 1984

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Note 84-53	2. GOVT ACCESSION NO.: AD-A138 223
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): The Validity of Performance Scores Adjusted for Environmental Conditions	5. TYPE OF REPORT & PERIOD COVERED:
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Laurel W. Oliver and Cecil D. Johnson	8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: US Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q263731A776
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ DA Deputy Chief of Staff for Personnel Human Resources Development Directorate WASH DC 20310	12. REPORT DATE: February 1984
	13. NUMBER OF PAGES: 16
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

Presented at the meeting of the American Psychological Association, Montreal, September 1980.

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

reliability of performance ratings
validity of performance ratings
field task performance
effect of environmental conditions on performance ratings

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

The purpose of this research was to investigate the validity of a measure of field task performance adjusted for environmental conditions. The difference between the correlation of the "adjusted" (corrected for adverse environmental conditions such as inclement weather) performance scores and daily ratings of overall job performance and the correlation of the "unadjusted" (raw score) performance scores with the daily ratings approached significance for the female sample. This finding is seen as encouraging for the further development of the correction procedure.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FOREWORD

The Leadership and Management Technical Area of the U.S. Army Research Institute for the Behavioral and Social Sciences is conducting research on organizational cohesion. The following report describes research related to the establishment and maintenance of unit cohesion.

This report describes a secondary analysis of data previously collected under Army Project 2Q263731A776. This in-house research was carried out under Army Project 2Q263731A792, Personnel Management, "Individual and Group Effectiveness," FY 80 Work Program.

THE VALIDITY OF PERFORMANCE RATINGS ADJUSTED FOR ENVIRONMENTAL CONDITIONS

BRIEF

Requirement:

Maintaining Army readiness requires valid and reliable assessments of soldier performance. This report investigates a method to increase the accuracy of performance evaluations by adjusting for the effects of environmental factors.

Procedure:

Soldiers were rated on various Soldiers' Manual tasks by officer evaluators during an extended field training exercise. These "raw" score ratings were adjusted to take into account the effects of seven different environmental conditions (e.g., climate, visibility, etc.). The unadjusted and adjusted ratings were then compared with supervisors' daily performance ratings to determine if the adjustments increased the accuracy of the officers' evaluations. Before the hypothesis that ratings which were adjusted for environmental conditions were more accurate than unadjusted ratings could be tested, it was necessary to determine if the evaluating officers and supervisors were measuring the same construct. A significant correlation between the raters' scores would indicate they were.

Findings:

Performance ratings of females by the two types of raters were significantly correlated. Hence, the hypothesis of increasing the validity (accuracy) of the ratings by adjusting them could be tested. It was found that the correlation of the adjusted ratings with supervisors' ratings was higher than the correlation of unadjusted ratings with supervisors' ratings. This difference between the two correlations approached statistical significance. In addition, adjusting the ratings made them significantly more reliable than the unadjusted ratings. Performance ratings of the males by the two types of raters were not significantly correlated, which precluded further analysis for this group.

Utilization of Findings

The technique described in this report shows promise for increasing the reliability and validity of performance ratings made in situations in which environmental conditions may affect performance ratings. More research is needed to develop the technique.

THE VALIDITY OF PERFORMANCE RATINGS ADJUSTED FOR ENVIRONMENTAL CONDITIONS

CONTENTS

	Page
INTRODUCTION	1
Background	1
Problem	2
METHOD	2
Subjects	2
Instruments	2
Data Collection	3
Data Analysis	4
RESULTS	4
DISCUSSION	5
Table 1. Correlations of Daily Performance Ratings with Unadjusted and Adjusted Individual Event Ratings for Males and Females	8
APPENDIX A. EXAMPLES OF INSTRUMENTS	9

THE VALIDITY OF PERFORMANCE RATINGS ADJUSTED FOR ENVIRONMENTAL CONDITIONS[1]

INTRODUCTION

Background

Rating scales have long been used to assess performance. When the performance being measured takes place in a location where conditions may vary widely, valid performance ratings become problematic. A soldier's performance on a task in an Army field environment, for example, may be viewed as due to his/her personal characteristics (such as ability, achievement, and motivation) plus an effect of the environmental conditions which surround the measurement.[2] Two recent research efforts by the Army Research Institute have focused on performance during field training exercises. The first of these, MAX WAC (Army Research Institute, 1977), was concerned with the effect of the proportion of women on the accomplishment of a unit's mission as measured by scores on ARTEP (Army Training and Evaluation Program) tasks. REF WAC, the second research effort (Johnson, Cory, Day, & Oliver, 1978), was primarily concerned with the comparative performance of men and women. In the earlier MAX WAC research, ARTEP performance measures were relatively imprecise for the MAX WAC Army officer evaluator used only a 3-point scale (1 = unsatisfactory; 2 = satisfactory; 3 = outstanding) to assess performance on ARTEP tasks such as erecting a hospital tent or serving as a convoy during night maneuvers. The evaluators used a very subjective procedure in arriving at their final ratings. For example, in evaluating the performance of a transportation unit moving supplies in a rainstorm, the evaluator might take into account the inclement weather in rating the quality of the unit's performance.

The REF WAC research built upon the MAX WAC experience and attempted to make more precise and objective the measures used in assessing individual performance on Soldier's Manual tasks such as investigating a traffic accident. For each individual event (task) to be evaluated, a set of criteria was constructed (e.g., "identifies and secures statements from drivers and witnesses"). Each criterion was rated as satisfactory or unsatisfactory, and a final rating was made on a 7-point scale. In addition, ratings of the effects of seven environmental conditions (such as weather or leadership) were averaged to obtain an environment score. The raw ("unadjusted") score and the environment score were summed to obtain an "adjusted" score. Thus, both the MAX WAC and REF WAC research attempted to correct for error caused by environmental conditions surrounding the measurement. The REF WAC measures, profiting from the MAX WAC research, were the more refined and objective measures.

---

[1]A version of this paper was presented at the meeting of the American Psychological Association, Montreal, September 1980. The authors wish to extend their appreciation to Dr. John J. Mellinger and Mr. Sidney A. Sachs for their assistance with the statistical analysis.and to Dr. Melvin J. Kimmel for his helpful comments.

[2]Plus, of course, measurement error. This paper addresses the problem of nonrandom bias caused by environmental conditions rather than traditional measurement error.

Purpose of Research

Although the REF WAC measures improved upon those developed for the earlier MAX WAC research, it was still not known to what extent the adjusted ratings used in REF WAC reduced error and thus more closely represented the individual's true level of performance. The purpose of the present research was to investigate the validity of the adjusted ratings. If the adjusted ratings were a more accurate assessment of an individual's performance, the adjusted ratings should correspond more closely to another measure of individual performance (daily performance ratings by the soldiers' supervisors) than would the unadjusted or raw score ratings. Accordingly, the following hypothesis was tested for males and for females:

> The correlation of adjusted individual event ratings with supervisors' daily performance ratings is significantly higher than is the correlation of unadjusted individual event ratings with supervisors' daily performance ratings.

METHOD

The instruments, subjects, and procedures used in the REF WAC research project are described in detail in Johnson et al., 1978. Brief descriptions of the methodological aspects which pertain specifically to the research reported here are given below.

Subjects

The sample for this research consisted of 151 enlisted personnel (97 women and 54 men) in 22 Army companies participating in an extended (11-day) field training exercise in Europe. The sample constituted a subset of a larger group of male and female service members who were the subjects of the REF WAC research described in Johnson et al. (1978). The larger group consisted of a female cohort (N = 200), which included all women in the 22 companies and a male cohort (N = 196) containing men who had been matched with the women on the basis of company, paygrade, length of service, military occupational specialty (Army job), and intelligence. Due to transfers of personnel to other units and to the fact that members of the male cohort were frequently not available for performance ratings on individual tasks, complete data were available for only 97 women and 54 men.

Instruments

Individual Performance Score Sheet (Individual Event Rating Module). This instrument was used by officer evaluators to rate individuals on individual events, which were specific Soldier's Manual tasks. The tasks selected were those which the researchers felt would be typical of a given type of unit and which were expected to occur with some frequency during the field training exercise (such as "Investigate Traffic Accidents" for Military Police units). The rating form contained scoring criteria that were equivalent, where possible, to those provided by the Training and Doctrine Command (TRADOC) for the Soldier's Manual task. Criteria for the task were enumerated on the form (e.g., "Measures and records skidmarks and final position of vehicles"), and the evaluator was

instructed to score each criterion as satisfactory, unsatisfactory, or not applicable. After considering the scores on all criteria, the evaluator then produced a performance rating on a 7-point scale. This scale incorporated key words and phrases from the operational Enlisted Evaluation Report (DA Form 2155-55), ranging from "Performed all tasks in a superior manner..." (rated 7) to "Performed all tasks in an inferior manner..." (rated 1).

In addition to the performance rating an "environment score" was also recorded. The environment score was the evaluator's estimate of the degree to which environmental conditions had affected the individual's performance. Each of seven environmental conditions (weather, terrain, equipment, night, leadership, instructions, and activity level) was rated on a three-point scale. The three points of this scale were identified as: optimal conditions (0), mildly adverse conditions (1), and severely adverse conditions (2). After considering his/her ratings of the seven environmental conditions, the evaluator recorded an overall environment score which reflected his/her estimate of the combined effect of the seven conditions on a 7-point scale ranging from 0 to 3 with half-point increments. The two scores just described (the raw score performance rating and the environment score) were then summed. Hence, the "unadjusted" score was the raw score performance rating, and the "adjusted" score was the raw score corrected for environmental conditions. (See Appendix A, page A-1, for an example of an individual event rating module.) The correlation of the unadjusted ratings and the adjusted ratings, based on 151 subjects (97 males and 54 females), was .90.

The reliability of the individual event ratings was also investigated. Two or more scores were available for 175 subjects (89 males and 97 females), for whom the mean number of ratings was 4.2. Since many subjects had more than two ratings and there were unequal numbers of ratings per subject, the intraclass correlation (Ebel, 1967) was used. The intraclass correlation was .44 for the unadjusted ratings and .57 for the adjusted ratings. A comparison of these intraclass correlations (Fisher, 1950) revealed that they differed significantly ($p < .05$), thus demonstrating that adjusting the performance ratings significantly increased the reliability of the ratings.

Daily Record of Work Availability and Performance (Schedule 4). This form was used by non-commissioned officer (NCO) data collectors to record supervisors' ratings of the performance of the male and female subjects. These daily performance ratings were on the same 7-point scale that was used for the individual event ratings described above and were measures of the subject's overall performance on the preceding day. The ratings were collected daily from the individual's supervisor.

To assess the reliability of the supervisors' daily performance ratings, a split-half procedure was used. Daily ratings were available for 278 subjects (141 from the male cohort, and 137 from the female cohort). As the correlation between the odd-day and the even-day ratings was .93 for this group, one can conclude that these ratings were highly reliable.

Data Collection

Individual event ratings. The performance of individuals on the Soldier's Manual tasks was rated by the officer evaluators, who used the Individual Performance Score Sheet described above. An effort was made to observe each of

the female soldiers on a variety of tasks. However, the evaluators had no control over the type of tasks performed or the frequency of task occurrence. After a woman was observed and rated, a man from the same company and as similar to her as possible in age and rank was rated on the same task.

Supervisors' daily performance ratings. The Schedule 4 (Daily Record of Work Availability and Performance) data were collected by noncommissioned officers (NCOs) assigned to each company. Each day during the field training exercise, the NCO data collectors obtained performance ratings for each male and female subject in their assigned units. These ratings were of the subject's overall performance on the preceding day.

Data Analysis

Before a meaningful test could be made of the hypothesis, it was necessary to determine if the two types of performance ratings (individual event ratings and supervisors' daily performance ratings) did in fact measure the same construct--namely, an individual's job performance during the field training exercise. Accordingly, the supervisors' daily ratings were correlated with (1) the unadjusted individual event ratings and (2) the adjusted individual event ratings for the male and female groups. A correlation which differed significantly from zero provided evidence of construct validity--i.e., both types of performance measures (individual event ratings and supervisors' daily ratings) appeared to have significant variance in common. In such a case, it was reasonable to continue the analysis and test the hypothesis by assessing the significance of the difference between the correlations of (1) unadjusted individual ratings and supervisors' daily ratings and (2) adjusted individual ratings and supervisors' daily ratings by using Hotelling's t test for the significance of the difference between nonindependent r's (Edwards, 1963). If the latter correlation should be higher than the former, the findings would suggest that adjusting the ratings makes them more accurate--i.e., increases their validity.

Results

The correlation of unadjusted individual ratings and supervisors' daily ratings (.32) and the correlation of adjusted individual ratings and supervisor's daily ratings (.40) were both significantly different from zero at the .001 level for the female sample. For males, however, neither the correlation between unadjusted individual task ratings and supervisors' daily ratings (.01) nor the correlation between adjusted individual task ratings and supervisors' daily ratings (-.09) varied significantly from zero.

In view of these findings, the hypothesis was tested only for the female group using a two-tailed t test for nonindependent r's (Edwards, 1963). The t value for the difference between the correlation of unadjusted individual event ratings and daily supervisors' ratings and the correlation of adjusted individual event ratings and daily supervisors' ratings was 1.90. This finding approached but did not attain significance ($t_{.05} = 1.99$, $df = 94$).[4] Table 1 summarizes these results.

---

[3] It would be possible to argue for a one-tailed test, although the authors chose not to do so. The difference between the correlations would be significant at the .05 level if a one-tailed test were used.

Discussion

The limitations of the data collection procedure should be kept in mind in considering the results of this research. The noninterference constraint, for example, limited the opportunities for observing individual events to such an extent that only 70 of the 113 different modules prepared were actually used. Observations of individual events were thus largely a matter of chance, with a low probability of observing the paired male. In addition, mean scores for individuals were sometimes based upon only one or two observations.

However, the findings show that the procedure for adjusting the raw score ratings had positive consequences for both the reliability and validity of the ratings. The reliability, for example, increased significantly ($p < .05$) as a result of applying the correction procedure.

The results reported here also provided evidence of construct validity for the two types of performance ratings used in this research. Correlations were significant between the unadjusted individual event ratings and supervisors' daily ratings and between the adjusted individual event ratings and supervisors' daily ratings. However, this finding applied only to the female subjects. These correlations were not significant for the male group.

This evidence of construct validity is of particular interest because the two types of performance ratings were very different. The individual event rating dealt with a specific task, while the supervisors' ratings were global in nature. Further, the individual ratings were made by an officer observer not connected with the soldiers' unit, whereas the supervisors' ratings were made by someone who knew the soldier well.

The findings also suggest that adjusting raw scores for environmental conditions results in a more valid measure of performance. For women, the difference between the correlation of the adjusted ratings with supervisors' ratings and the correlations of the unadjusted ratings with those ratings approached significance. Therefore, it appears that the procedure for adjusting the raw scores for environmental conditions had positive consequences for the validity of the performance ratings. While not definitive, these findings are viewed as encouraging for the further development of the correction procedure.

Since the positive results concerning validity were found only for the female sample, the question arises as to why this outcome occurred. The focus of the data collection, as perceived by the officers and NCOs charged with collecting data, was on female performance. It appeared that officers and

NCOs who collected data made a special effort to follow the specified procedures for women and were somewhat more casual about the collection of parallel male data because they believed the latter to be of lesser importance.[4] Improved data collection procedures would minimize such problems, and perhaps confirm the hypotheses for male as well as female subjects.

[4] In fact, at one point about halfway through the field training exercise, it was necessary to instruct data collectors to observe more all-male groups as most of the group event data being collected were on groups containing one or more women. Although the group event ratings are not considered in this paper, the incident does indicate the emphasis the data collectors placed on observing the performance of women.

References

Army Research Institute. Women content in units force development test (MAX WAC). (ARI Special Report S-6). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, 1977.

Brichtson, C. A., Ciaverelli, A. P., & Wulfeck, J. W. Operational measures of aircraft carrier landing system performance. Human Factors, 1969, 11, 281-290.

Ebel, R. L., Estimation of the reliability of ratings. In W. A. Mehrens & R. L. Ebel (Eds.), Principles of Educational Psychological Measurement. Chicago: Rand McNally, 1967.

Edwards, A. L. Experimental design in psychological research. (Rev. Ed.) New York: Holt, Rinehart and Winston, 1963.

Fisher, R. A. Statistical methods for research workers. (11th Ed.) New York: Hafner, 1950.

Johnson, C. D., Cory, B. H., Day, R. D., & Oliver, L. W. Women content in the Army - REFORGER 77. (ARI Special Report S-7). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, 1978.

Table 1

Correlations of Daily Performance Ratings with Unadjusted and Adjusted Individual Event Ratings for Males and Females

Group	Individual Event Ratings		t [a]
	Unadjusted	Adjusted	
Males (n = 54)	.08	-.01	.40 (n.s.)
Females (n = 97)	.32***	.40***	1.90 (n.s.)

[a] t test for the significance of the difference between nonindependent r's (two-tailed test)

***$p < .001$ that correlation differs significantly from 0.

APPENDIX A

Examples of Instruments

INSTRUCTIONS
(INDIVIDUAL EVENT RATING MODULE)

*EVENT TITLE: Descriptive title of event

*TOE AUTHORIZED JOB TITLE: As extracted from appropriate line in authorization document.

*TEAM CODE SEQUENCE NUMBER: Includes team designation and numerical team sequence number as assigned by the team chief e.g., MP01, TC11, SC13, MC03, MT16.

NAME - FEMALE: Name and rank of subject.

NAME - COHORT: Same as above. A cohort must be tracked with each female.

UNIT: Designation of company to include battalion headquarters if lettered company.

DTG: Date and time of observation in six numerical digits with the date first-e.g.141305 Aug 77

OBSERVATION TIME LENGTH: The time which transpired during the observation.

TIME SINCE LAST REST/WORK BREAK: The amount of time subject has been on duty or since shift began.

MOS: Primary, secondary and duty MOS in that order. Do not interfere with unit. Obtain from NCO data collector and fill in later.

MOS EXPERIENCE: Time in duty MOS. Obtain from NCO data collector.

TIS: Length of active federal military service. Obtain from NCO data collector.

ENVIRONMENT: Score all items as:
O - Optimal conditions
M - Mildly adverse; some negative effect
S - Severely adverse; substantial negative effect
NA - Environment not applicable

WEATHER: Did weather conditions (precipitation etc.) inhibit individual in completing the task?
TERRAIN: Did the topography (vegetation, road network, etc) inhibit individual in completing the task?
EQUIPMENT: Did equipment shortage, unserviceability, etc, inhibit individual in completing the task?
NIGHT: Did natural light have any effect on individual performance?

LEADERSHIP: If supervision is applicable to event, did its quality affect performance?

INSTRUCTIONS: Was individual properly briefed?

ACTIVITY LEVEL: Was individual unduly stressed by the frequency of events or external factors? Do not confuse with fatigue.

*SCORING CRITERIA: Score each event/sub-event as; S - Satisfactory U - Unsatisfactory NA - Not applicable

PERFORMANCE SCORE: (1-7):
7 - Performed all tasks in a superior manner; equivalent to the performance of an outstanding soldier.
6 - Performed most tasks in a superior manner, all others at minimum standards; equivalent to the performance of a superior soldier.
5 - Performed some tasks in a superior manner, all others at minimum standards; equivalent to the performance of an excellent soldier.
4 - Performed all tasks at minimum standards; equivalent to the performance of an average soldier.
3 - Performed most tasks at minimum standards; some tasks were failed; equivalent to the performance of a marginal soldier.
2 - Performed a few tasks at minimum standards but most were failed; equivalent to the performance of an unsatisfactory soldier.
1 - Performed all tasks so inferior as to question MOS qualification of individual.

ENVIRONMENT SCORE (0-3): Average letter scores from environmental section above and apply to the following scale, place a number at the bottom of each column in space provided:

0	0.5	1.0	1.5	2.0	2.5	3.0
O			M			S

OBSERVER: Name of observer

EVENT SHEET NUMBER: Sheet sequence number if observer completes more than one sheet for same individual event.

To be filled in in advance.

MODERATOR SCORE _____ (Maximum 3 pts)

SP 77-2b

PROJECT REF VAC
INDIVIDUAL PERFORMANCE SCORE SHEET
(Three Observations Minimum)
INDIVIDUAL EVENT RATING MODULE

EVENT TITLE Investigate Traffic Accidents
TOE AUTHORIZED JOB TITLE Military Policemen
TEAM CODE SEQUENCE NUMBER MP01

REFERENCE DATA:								
Name-Female								
Name-Cohort								
Unit								
DTG								
Observation Time Length								
Time Since Last Rest/Work Break								
MOS								
MOS Experience (Time)								
TIS								
ENVIRONMENT:								
Weather								
Terrain								
Equipment								
Night								
Leadership								
Instructions								
Activity Level								
SCORING CRITERIA:								
1. Assesses situation and insures continued flow of traffic.								
2. Renders first aid and assists with the evacuation of the injured.								
3. Protects the accident scene until essential facts are gathered.								
4. Note positions of fatalities and outline in chalk.								
5. Aid in clearing obstructions.								
6. Assists with evacuation of damaged vehicles.								
7. Requests additional MP support, (requests emergency vehicles and fire and ambulance).								
8. Secures and protects evidence by cordoning off area.								
9. Identifies and secures statements from drivers and witnesses.								
10. Measures and records skidmarks and final position of vehicles.								
11. Prepares traffic accident report (DA Form 3946) to include basic sketch of scene.								
12. Reestablished normal flow of traffic after accident scene has been cleared of vehicles and debris.								
PERFORMANCE SCORE (1-7):								
ENVIRONMENT SCORE (0-3):								

OBSERVER: _____

EVENT SHEET NUMBER: _____

Data collector - signature

SCHEDULE 4
REVORGEN 77 - ARI
Daily Record of Work Availability and Performance

Date _____ day month year
Company _____

Column: 1 Name Last, First, Initial	2 Sex	3 Pay Grade	4 Daily Duty Assignments: Total hrs. Assigned	5 Daily Duty Assignments: Total hrs. not Available	6 Reason for Non-availability or Impairment (code)	7 Source	8 Rating of Performance (code)	9 Source	10 COMMENT
1.									
2.									
3.									
4.									
5.									
6.									
7.									
8.									
9.									
10.									

A-3

10 Aug 77

PT 5167t