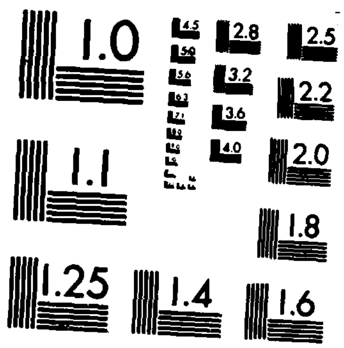
AD-A174 843 A COMPARISON OF TWO TASK RATING SCALES OF PHYSICAL DEMAND(U) AUSTRALIAN MILITARY FORCES CANBERRA PSYCHOLOGICAL RESEARCH UNIT (1) R S COLLYER AUG 86 1/1

UNCLASSIFIED RR-3/86 F/G 5/9 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

RR 3/86

AR-005-004

AD-A174 843

A COMPARISON OF TWO TASK RATING SCALES OF PHYSICAL DEMAND

BY

MAJOR R.S. COLLYER

1st PSYCHOLOGICAL RESEARCH UNIT

86 11 12 070

Research Report 3/86

A COMPARISON OF TWO TASK RATING SCALES OF PHYSICAL DEMAND

by

Major Robert S. Collyer

This Directorate of Psychology publication has been prepared by the 1st Psychological Research Unit and is authorized for issue by DPSYCH-A.

LTCOL P. N. DRAKE-BROCKMAN
Commanding Officer
1st Psychological Research Unit
NBH 3 - 44
Northbourne House
TURNER ACT 2601

Paper presented at the 21st Annual Conference of the Australian Psychological Society, James Cook University, Townsville, QLD. 25-29 August, 1986.

TABLE OF CONTENTS

LIST OF TABLES

III

Abstract

Task analysis to specify the physical demands of work is expensive in terms of manpower skills, equipment, and time. Job analysis methods to identify tasks that should be subject to detailed task analysis are an important part of any strategy to specify physical demands criteria for jobs. This paper reports research comparing two sets of task inventory rating scales used to identify important, physically demanding tasks. The physical demand scales are Perceived Physical Effort, which is part of Physical Abilities Analysis, and Physical Strength and Endurance, a scale developed by the US Air Force. Two scales assessing Consequences of Inadequate Performance were also used. The scales were administered to large samples from two Australian Army employment groups. Scale reliabilities and task rank order intercorrelations are reported and implications for the scales' future use are discussed. (Australia).

This research arose from the Australian Army's need to specify the physical demands of Army jobs. A rigorous methodology was required, but it had to take into consideration the need for (a) economy of resources in the setting of criteria, (b) the limited time available for the testing of personnel as part of the process of allocating recruits to trades, and (c) the availability of physical capacity test(s) related to job criteria.

A literature search (Collyer & James, 1985) indicated that a range of approaches have been employed by other researchers, but as Campion (1983) points out, most approaches can be categorized as: (a) measures of metabolic requirements, (b) measures of strength requirements, or (c) measures of multiple physical abilities.

Measurement of metabolic requirements involves "physiological" measures like those of oxygen uptake or heart rate during task performance. Oxygen uptake measures can be made by collecting and analysing expired oxygen using a Douglas Bag (Astrand & Rodahl, 1977), or Kofranyi-Michaelis gas meter (Consolazio, 1971). Measures of heart rate are an indirect method of predicting oxygen consumption because they are linearly related within an individual (Astrand & Rhyming, 1954). Although these metabolic methods are the most accurate, they are cumbersome to administer and require expensive instrumentation and trained personnel. These latter three characteristics are undesirable in a method for large scale implementation.

Measures of strength requirements rely on assessments of weights lifted, heights to which they are lifted, and distances carried. A variety of methods are available for the assessment of human strength, but they all utilize one of three muscle contractions: (a) isometric, (b) isotonic, or (c) isokinetic. Isometric strength is relatively easy to measure accurately using practical standardized methods, and some researchers (e.g., Chaffin, 1975) recommend its use, although care must be exercised when doing so since isometric (also known as static) strength is not perfectly correlated with the other strengths. Research (Ayoub, et al. 1982; Hogan, Ogden, Gebhardt, & Fleishman, 1979; Sharp, et al. 1980) has shown that approximately 90 percent of tasks that are physically demanding are so because of weights lifted and carried. The focus of these findings on strength, combined with the practicalities of measuring isometric strength, offered hope of a practical, economic method for Army use.

Measures of multiple physical abilities involve assessment of such characteristics as strength, stamina, body flexibility, and balance. Physical Abilities Analysis (Fleishman & Hogan, 1978; Myers, Gebhardt, & Fleishman, 1980a; 1980b) is a major example of this technique. It involves experienced job personnel making judgments about task demands using nine specific fitness scales and one general effort scale. The scales have been empirically derived from correlational studies and require little training or equipment to administer. The economy of using these rating scales was attractive, but selection tests to assess all the abilities measured by Physical Abilities Analysis were expected to be more complex than would be the case if measures of strength alone were utilized.

Each method has advantages and disadvantages, but each is a form of task analysis that is designed to be applied to specific tasks. Most jobs consist of a multitude of tasks, and most analyses are conducted to develop selection criteria to be applied to job types rather than to individual jobs or positions. These job types often involve hundreds or even thousands of tasks. Regardless of which form of task analysis is employed, job analysis to identify tasks which should be analysed in detail is an important preliminary to the specification of physical demands criteria, although the form of job analysis must take the type of task analysis into account. This paper discusses a method of job analysis that has been used for this purpose.

Methods of job analysis that include assessment of physical demands are available. Examples are the Position Analysis Questionnaire (McCormick, Jeanneret and Mecham, 1969), Functional Job Analysis (Fine and Wiley, 1971) and Physical Abilities Analysis (Fleishman and Hogan, 1978). Another job analysis method that provides detailed information is the task inventory survey method paired with the Comprehensive Occupational Data Analysis Programs analysis package (TI/CODAP).

The Australian Army has used this method of job analysis since 1975. TI/CODAP was considered to meet two requirements of the research, those of rigorous methodology and economy. It provides an economical way to obtain job task data from large numbers of workers in diverse locations. The information, obtained from job incumbents and/or supervisors, is detailed, quantifiable data that can be manipulated by CODAP to provide a wide range of group job descriptions. Because the data are quantifiable they can be validated, and the general methodology is supported by a large body of sound research (see Christal, 1974, for examples).

The TI/CODAP job analysis uses survey questionnaires administered by mail. The main components of these questionnaires are a task inventory, listing all significant tasks in the employments surveyed, and task rating scales. Experienced supervisors use the scales to rate task characteristics such as the amount of training emphasis that should be given in formal training, or the consequences of inadequate performance of a task. Task rating scales such as these have been researched and used extensively overseas (e.g., Christal, 1974; Jansen, 1982) and in Australia (e.g., limited distribution occupational analysis reports within the Department of Defence), but there was no Australian research using scales of physical demand. Prior experience had shown that task rating scales used successfully overseas did not necessarily work for the Australian Army (e.g., the Training Emphasis scale). This was usually because of the diverse nature of Australian employments. It was decided to select and pilot scales that seemed to identify physically demanding tasks to determine their appropriateness to the Australian situation.

The literature search had identified two main streams of research that offered promise of suitable task rating scales. The first was work originating from the US Air Force (who were the original developers of the TI/CODAP method). The second was Physical Abilities Analysis, already mentioned as a form of task analysis. Scales from these research areas were selected.

US Air force research reported by Gott and Alley (1980) showed that a ten point rating scale (0 to 9) could be used to elicit reliable ratings of physical strength and endurance. Reliability was assessed using the Lindquist (1953) method of intraclass correlation. This reliability is in itself a form of convergent validity, and other validity studies were underway, so the scale was selected for trial. The Physical Strength and Endurance (PSE) scale was modified to provide metric measures of weight and height, and to combine the two lower categories, thus making a nine-point scale anchored at each point (see Appendix 1). This modification was necessary because the version of CODAP on line in the Department of Defence was unable to distinguish between "X" and "0" as scale ratings. The "X" for "No knowledge of task requirement" was considered necessary to the research, and the distinction between the "0" and "1" scale point definitions was not vital to the research objectives.

Physical Abilities Analysis comprises ten scales. Nine of these scales assess specific abilities (Static Strength, Dynamic Strength, Explosive Strength, Trunk Strength, Stamina, Extent Flexibility, Gross Body Coordination, and Equilibrium). Whilst these scales have been shown to have good reliability and validity (Hogan, et al. 1979; Myers, et al. 1980a; 1980b), this same research has shown the tenth scale, Perceived Physical Effort (PPE), to have good reliability and validity (i.e., it correlates highly with the specific scales and with actual metabolic costs) in identifying physically demanding tasks. Use of the full ten scales for each Army employment group was seen as a major undertaking and one which would be better avoided if possible. Because of the good reports on the PPE scale, it was selected for research and comparison with the PSE one.

Because it is not a reasonable proposition to use tasks which are not important as criterion tasks by which job demands are specified, a measure of the consequences of inadequate performance of tasks was incorporated into the research. Two Consequences of Inadequate Performance scales were selected. Both scales assess the same task characteristics, but one (CIP-9) is a nine-point scale anchored at each point to pair with the nine-point PSE scale and the other (CIP-7) is a seven-point scale anchored at the mid-point and extremities to pair with the seven-point PPE scale.

A comprehensive job analysis as proposed, while being a relatively economical way to obtain job data from a large number of workers, is still time consuming in that it requires experienced workers to spend between two and three hours completing the questionnaire. Determination of optimum sample sizes can provide substantial savings in time whilst retaining the reliability and validity of the results. This research incorporated a number of checks to aid in determining optimum sample size. Samples were from employments selected to provide a substantial range of variables representative of normal Army employments, and were expected to provide a rigorous evaluation by comparison of reliability coefficients and rank order correlations of task mean ratings for sub-samples. Rank order correlations were used because sets of critical, demanding tasks were to be selected for task analysis. The scales were not required to specify actual task demands in terms of weights, heights, distances, or aerobic demand, even though one scale used actual weights and heights as scale-point definitions.

This paper reports research undertaken to evaluate the two pairs of task rating scales, PSE/CIP-9 and PPE/CIP-7, as scales that can be used to identify important, physically demanding tasks for detailed task analysis. In addition to assessing the suitability of the scales, the research also sought to identify optimum sample sizes for the job analysis. Although the job analysis method used was TI/CODAP, which is not commonly available, the rating scales and procedures are ones that can be used with other analysis packages.

Method

Samples

Developing samples for surveys of this type is a matter of experience and judgment. The aim is to select samples that are representative of the range of work performed and the conditions under which it is performed. To evaluate a level of confidence in the results, it is necessary to quantify the agreement between raters. The Lindquist (1953) method of assessing interrater agreement was used. Experience in Australia and research overseas (e.g., Jansen, 1982) has shown that, depending on the homogeneity of the employment, a suitable level of interrater agreement can be obtained using 30 to 50 raters. When the employments are heterogeneous (i.e., have a large diversity of job types), this should be interpreted as 30 to 50 raters per major analysis variable, so agreement can be evaluated for sub-samples in case different rating policies exist. Because many Australian Army employments are heterogeneous by comparison to US Forces' ones (where the selected scales were developed), it was necessary to select samples that would enable assessment of the scales with heterogeneous employment groups

The sampling strategy employed was a purposive one. It was designed to obtain as many experienced raters as possible up to at least 40 in each identified major analysis variable (although in some cases it was not possible to obtain 40 raters). In addition to these considerations, the sample selected was larger than normally considered necessary for good quality information so that the data could still be used if unforeseen variables were later identified. Two employment groups were selected for sampling, the vehicle mechanics and the driving trades.

Vehicle mechanics. A target sample of 314 vehicle mechanics was selected for survey. This employment was chosen because it is a technical employment with a wide range of work and work conditions. Vehicle mechanics do all types of mechanical repairs and maintenance on vehicles ranging from light vehicles (e.g., staff cars and motor cycles) to heavy tanks, tank transporters, earth movers, and dockside cranes. There are two main types of repair work, base repair, where major and minor repairs are conducted on all types of vehicles in base workshops, and field repair, where major or minor repairs are conducted on all types of vehicles using equipment at hand in the field or in mobile workshops. Field repair work is subject to the vagaries of both weather and combat movement. These two types of repair work provide two major variables, with base repair units being generalized as Non Field Force Command units and field repair units as Field Force Command ones. Previous occupational analysis studies have shown this employment group to be homogeneous with respect to tasks performed (Healy, 1984) but the varying equipment and work conditions were expected to provide more variability for this type of study.

The sample included all vehicle mechanic Sergeants, Staff Sergeants and Warrant Officers, and some experienced Corporals. The two pairs of survey scales were allocated randomly among the sample. The returned questionnaires (the actual sample) represented 87.6 percent of the intended sample, and there were 145 PSE/CIP-9 questionnaires and 130 PPE/CIP-7 ones.

Driving trades. From a population of 1951 drivers, 950 were selected for survey (the intended sample). The driver's employment group includes operators of specialist vehicles and transport supervisors, as well as drivers of trucks and staff cars across a number of Corps. In some Corps there were few drivers, in others many. Drivers from five Corps (Artillery, Infantry, Signals, Engineers and Transport) were selected for sampling because they were considered to be representative of drivers' work. This was seen as a heterogeneous group and one where it was known that some drivers had difficulty in coping with the physical demands of some tasks.

The sample was selected to obtain representation of trade and Corps and also of unit type (i.e., Field Force Command or Non Field Force Command). Corps and unit type were the major variables. The two pairs of rating scales were randomly allocated among the sample. A total of 365 PSE/CIP-9 questionnaires and 349 PPE/CIP-7 ones were processed (the actual sample). These represented 77.2 percent and 73.2 percent of the target sample respectively, and 36.6 percent of the Driving Trades population.

Panel raters. Because job analysis surveys are time consuming to administer and take respondents from their work, checks were incorporated into the research design to test for ways of reducing the demand on time and resources. The task inventories were developed by panels of experienced workers (described in the next section) and one check was to arrange for these panel members to provide task ratings using the PSE scale for both employment groups, and the CIP-9 scale for one group. It was not possible to test the PPE/CIP-7 scales in this way because of a shortage of time at the completion of the inventory development panels. These panel members, who were considered to be knowledgable raters, had just completed four to five days analysing the employments and their ratings were to be compared to the major sample ratings obtained later in the research. These ratings, for the vehicle mechanics, were also compared to the panel members' own ratings (using the same scales) obtained during the major survey (test-retest reliability). Seven panel members provided ratings on PSE and CIP-9 for the vehicle mechanics questionnaire both as experienced panel raters and as raters during the main survey. Fourteen panel members provided ratings as experienced panel raters for the driving trades questionnaire.

Apparatus

Survey questionnaires. Four survey questionnaires were constructed, with one task inventory and set of background information questions and two pairs of task rating scales per employment group. Previous job analyses had provided job task data which were held as part of the Military Employments Data Bank. Existing task inventories from this data bank were used as the basis for experienced panels. Panel members, representing different areas of the employment, updated and restructured the inventories under the guidance of a trained inventory developer. Inventory development procedures were normal Army occupational analysis ones as detailed in Standing Operating Procedures for the Military Employments Research and Information Team (1982). Equivalent results can be obtained by following the procedures contained in Archer and Fruchter (1963). Background information questions, survey instructions, and rating scales were added later. The final questionnaire for the vehicle mechanics contained 501 tasks and 16 background information questions. The driving trades questionnaire contained 349 tasks and 17 background information questions.

Rating scales. Two pairs of rating scales were employed. Physical Strength and Endurance paired with the nine-point Consequences of Inadequate Performance scale, and Perceived Physical Effort paired with the seven-point Consequences of Inadequate Performance scale.

Physical Strength and Endurance (PSE) is defined as involving significant use of the "large" muscle groups in the arms, back, or legs. These include requirements for lifting, lowering, or carrying heavy or cumbersome objects, pushing or pulling, torquing, or any other demand for frequent or continuous exertion or muscular effort. Raters were told to make their ratings on the basis of (a) the most demanding aspects of each task, (b) the level of demand placed on a single individual performing the task, and (c) the level of demand required by the complete task from start to finish (Appendix 1-2). The scale has nine points, with anchor points described in terms of weights, heights, or equivalent muscular effort. Raters are given the opportunity to indicate with an "X" if they feel they do not know enough about the task requirements to provide a rating.

Perceived Physical Effort (PPE) is defined as the degree of physical exertion experienced in performing a single task or series of tasks. The scale has seven points with example tasks at the lower end, near the middle, and towards the upper end (Appendix 2). These example tasks are located on the scale according to the relative metabolic cost of their performance. Raters were asked to rate how much much effort it takes to perform each task.

Consequences of Inadequate Performance (CIP) is a measure of the seriousness of probable consequences resulting from a task not being performed adequately. It is defined in terms of probable injury or death, failure to accomplish a critical mission, wasted supplies, damaged equipment, and wasted hours of work. Two CIP scales were selected. Each scale measures the same characteristic. The differences between them are that one is a nine-point scale (CIP-9) anchored at each point, to pair with the nine-point PSE scale (see Appendix 3), the other is a seven-point scale (CIP-7) anchored at the mid-point and the extremities to pair with the seven-point PPE scale (see Appendix 4). The seven-point scale uses example tasks to help define scale-point anchors.

Procedure

Survey administration. The surveys were conducted by mail using nominated officers within units to administer them. Detailed instructions for completing the surveys were included in the questionnaire booklets and were supplemented by administrative instructions to the administering officers. Raters were instructed to rate all tasks using their total military experience, not just the situation in their present unit. After rating the tasks, respondents completed the background information questions and sealed the booklet in an envelope addressed to the project team. This was handed to the administering officer for return to Canberra. There was no requirement for booklets to be completed in the presence of the administering officer. These administrative arrangements were normal for Army occupational analysis surveys and most respondents had probably completed a survey booklet at some time in their career.

Results and Discussion

Scale Transformations

Since each survey respondent can have a personal assessment of what is "average" and rate according to that assessment when using the CIP scales, CIP-9 and CIP-7 were treated as relative scales. The PPE scale was also treated as a relative scale since it was judged that most raters would fix a frame of reference of "effort" within which to rate tasks. Ratings for these three scales were standardized to a mean of five and a standard deviation of one. This is a standard option available as part of the CODAP analysis system. Because the wording of the anchor points for the PSE scale made reference to specific weights and heights, it was judged to be an absolute scale and therefore unsuited for standardization.

Interrater Reliability.

Interrater reliability statistics used were based on the Lindquist (1953) method of intraclass correlation as applied using the CODAP REXALL program. Two indices of reliability are normally reported by REXALL, and a third has been calculated. The first, R_{11}, is the single rater reliability, which approximates the average of all possible pairwise correlations. The second, R_{kk}, is the reliability for a sample of k raters, which is the expected correlation between the set of observed sample task means and the task means of an hypothetical equivalent sample. If calculated R_{11} and R_{kk} values meet or exceed .20 and .90 respectively; they are interpreted as meaning that sufficient rater agreement exists to produce stable estimates of task mean values (see Jansen (1962) and Goody (1976) for more detailed discussions on the use of REXALL). Sample size is also an important consideration in assessing interrater reliability, especially in evaluating causes of poor agreement. Experience has shown that between 30 and 40 raters per analysis category can be expected to give reliable results. Because sub-sample sizes vary greatly in this research, the Spearman-Brown prophecy formula, which algebraically describes the relationship between sample size and interrater reliability, was used to calculate a standard R_{kk} for a sample size of k=50 raters.

Interrater agreement results are summarized for the vehicle mechanics sample in Table 1 and for the driving trades sample in Table 2. Comparison of R_{11} and R_{kk} results against the desirable criteria of .20 and .90 (Jansen, 1962) respectively indicate that results for the total rating set on all scales were very good for the vehicle mechanics, showing the close agreement between supervisors as to which tasks were important and which ones were physically demanding. The reliability results for the driving trades were also satisfactory, though the R_{11} values tended to be lower than those for the vehicle mechanics. These figures for the drivers reflect the more diverse nature of that employment group, but all values are satisfactory. It is also noteworthy that the R_{11} and R_{kk} values for the panel raters of both surveys showed good agreement, as did the values calculated for the vehicle mechanic panel members' ratings from the actual survey. It was not possible to collect survey data for the driving trades panel members during the actual survey.

Table 1

Interrater Reliability -- Vehicle Mechanics

CATEGORY	SCALE	R_{11}	R_{kk}	$R_{50\ 50}$	k
All raters	PSE	.52	.99	.98	134
	CIP-9	.41	.99	.97	134
	PPE	.62	.99	.99	122
	CIP-7	.45	.99	.98	122
FF Comd raters	PSE	.53	.98	.98	50
	CIP-9	.42	.97	.97	50
	PPE	.65	.99	.99	42
	CIP-7	.47	.97	.96	42
Non FF Comd raters	PSE	.52	.99	.98	83
	CIP-9	.40	.98	.97	84
	PPE	.60	.99	.99	80
	CIP-7	.44	.98	.97	80
Experienced panel	PSE	.56	.90	.98	7
	CIP-9	.46	.85	.98	7
Panel/survey	PSE	.67	.93	.99	7
	CIP-9	.48	.86	.98	7

Table 2

Interrater Reliability -- Driving Trades

CATEGORY	SCALES	R_{11}	R_{kk}	$R_{50\ 50}$	k
All raters	PSE	.36	.99	.97	261
	CIP-9	.26	.99	.95	260
	PPE	.42	.99	.97	260
	CIP-7	.27	.99	.95	260
FF Comd raters	PSE	.35	.99	.96	138
	CIP-7	.21	.97	.93	144
	PPE	.40	.99	.97	134
	CIP-7	.27	.98	.95	136
Non FF Comd raters	PSE	.37	.99	.97	124
	CIP-9	.28	.98	.95	126
	PPE	.44	.99	.97	125
	CIP-7	.27	.98	.95	125
Corps 1 raters	PSE	.38	.99	.97	145
	CIP-9	.28	.98	.95	146
	PPE	.45	.99	.98	144
	CIP-7	.28	.98	.95	144
Corps 2 raters	PSE	.39	.96	.97	34
	CIP-9	.26	.92	.95	34
	PPE	.39	.96	.97	34
	CIP-7	.21	.91	.93	36
Corps 3 raters	PSE	.35	.96	.97	41
	CIP-9	.22	.91	.93	38
	PPE	.38	.96	.97	38
	CIP-7	.23	.92	.94	39
Corps 4 raters	PSE	.31	.92	.96	26
	CIP-9	.25	.89	.94	26
	PPE	.40	.96	.97	32
	CIP-7	.33	.94	.96	33
Corps 5 raters	PSE	.33	.89	.96	17
	CIP-9	.28	.88	.95	18
	PPE	.36	.89	.97	12
	CIP-7	.26	.82	.95	13
Experienced panel	PSE	.61	.94	.99	14

The interrater reliability statistics indicate that personnel experienced in the employments surveyed can agree closely on which tasks are physically demanding and which ones are not. They can also agree which tasks are important and which ones are not when considered in relation to unit mission, safety, or loss/damage to stores. This agreement holds for all raters and for sub-groups by Corps or type of Command for each of the four scales. Use of the Spearman-Brown prophecy formula indicates that, should an occupational analysis survey be conducted using any of these scales, reliable results can be expected by surveying approximately 50 experienced trade personnel. This demonstrates a considerable saving in staff time, and in job incumbent time required to complete survey questionnaires. Time per person to complete these questionnaires was typically between two and three hours. The high interrater reliability results from the panel members ratings suggest that this set of task means may also be adequate for identifying tasks for task analysis. The reliability results for k=50 raters are consistent with previous reports (e.g., Jansen, 1982).

Scale Intercorrelations

Part of the survey aim was to compare how the pairs of rating scales selected important physically demanding tasks for task analysis. To make this comparison, the rank ordering of the tasks based on task mean ratings was assessed for PSE versus PPE and CIP-9 versus CIP-7. Tables 3 to 6 report the rank order intercorrelations for these comparisons. For the drivers, correlations for the ratings across the five Corps are summarized in Table 7. All correlations are significant (P<.01).

Table 3

PSE/PPE Rank Order Intercorrelations -- Vehicle Mechanics

	2	3	4	5	6	7	8
1. PSE all raters	.99	1.0	.94	.95	.96	.96	.95
2. PSE FF Comd	--	.98	.94	.96	.94	.94	.93
3. PSE Non FF Comd		--	.93	.94	.96	.96	.96
4. PSE panel raters			--	.95	.90	.90	.89
5. PSE panel/survey				--	.90	.90	.89
6. PPE all raters					--	1.0	1.0
7. PPE FF Comd						--	1.0
8. PPE Non FF Comd							--

Table 4

CIP-9/CIP-7 Rank Order Intercorrelations -- Vehicle Mechanics

	2	3	4	5	6	7	8
1. CIP-9 all raters	.98	.99	.75	.89	.97	.95	.97
2. CIP-9 FF Comd	--	.94	.81	.91	.97	.96	.95
3. CIP-9 Non FF Comd		--	.70	.85	.95	.92	.96
4. CIP-9 panel raters			--	.78	.78	.82	.76
5. CIP-9 panel/survey				--	.88	.86	.87
6. CIP-7 all raters					--	.98	.99
7. CIP-7 FF Comd						--	.95
8. CIP-7 Non FF Comd							--

Table 5

PSE/PPE Rank Order Intercorrelations -- Driving Traces

	2	3	4	5	6	7
1. PSE all raters	.82	1.0	.83	.93	.92	.94
2. PSE FF Comd	--	.99	.82	.92	.92	.93
3. PSE Non FF Comd		--	.83	.94	.93	.94
4. PSE panel raters			--	.72	.71	.74
5. PPE all raters				--	.99	.99
6. PPE FF Comd					--	.98
7. PPE Non FF Comd						--

Table 6

CIP-9/CIP-7 Rank Order Intercorrelations -- Driving Traces

	2	3	4	5	6
1. CIP-9 all raters	.99	.99	.97	.97	.97
2. CIP-9 FF Comd	--	.96	.97	.97	.95
3. CIP-9 Non FF Comd		--	.97	.96	.96
4. CIP-7 all raters			--	.99	.99
5. CIP-7 FF Comd				--	.97
6. CIP-7 Non FF Comd					--

Table 7

Range and Means for Rank Order Intercorrelations of Four Task Factors Across Five Corps

	MIN	MAX	MEAN
PSE	.78	.96	.86
PPE	.81	.94	.87
CIP-9	.79	.93	.86
CIP-7	.74	.94	.84

The scale intercorrelations show that the experienced raters agreed closely which tasks were physically demanding and which ones were not, whether they used the PSE scale or the PPE one. High agreement was also evident for the CIP-9 and CIP-7 scales. These results suggest that essentially the same job analysis results may be obtainable regardless of which pair of scales are used. This is discussed in more detail later in this paper. The high intercorrelations between the sub-samples selected by major variables (i.e., Corps and unit type) were expected (although not mandatory) once the high interrater reliability statistics had been calculated for the sub-samples. The high rank order correlations between the panel members' ratings and those of the main survey suggest that, under certain circumstances, the main survey may not be required. These circumstances are also discussed later.

Scale Validities

Although a high level of agreement exists between raters using both pairs of scales, there were some differences in the way the scales identified tasks, and in order to report these it is necessary to look briefly at the selection of tasks for analysis, and some results of that task analysis. This will give an indication of the validity of the physical demands scales for identifying tasks for task analysis.

Tasks were sorted in order of task mean ratings and categorized using the following decision rules: Firstly, those tasks with a mean rating greater than the grand (overall) mean plus one standard deviation on the physical demands scales were considered as potential tasks for task analysis; secondly, these tasks were categorized using the CIP mean ratings so that those tasks with means greater than the grand mean plus one standard deviation were included on a top priority task analysis list, and so on for lessor CIP categories.

Structured task analysis interviews using a format modified from Ayoub, et al. (1982) were conducted to identify precisely what made these important tasks physically demanding. Objects were identified and estimates of weights lifted, distances objects were moved, and forces (e.g., torques, push, pull) applied were obtained. Objects were also weighed and technical manuals checked for heights and torques. Estimates of frequency of task performance were also obtained. Part of these interviews involved seeking to identify physically demanding tasks which were not included in the list of tasks for task analysis. Although a few were proffered in the interviews, a check invariably showed that these had been included in the original job analysis and had not been on the task analysis list because mean ratings had been too low. No new tasks with significant physical demands were identified, indicating that the job analysis using both these physical demands scales and this method of selecting tasks had been quite adequate.

The choice of physical demands scale had several effects on the set of tasks identified by the task selection criteria. If the PPE scale was used, more tasks were considered to be physically demanding than by selecting with the PSE scale. But the PPE scale identified all significant tasks identified by the PSE one. Inspection of the tasks indicates that this appears to occur because raters using the PSE scale identify tasks largely on the basis of the weights (and possibly the height) specified in the scale anchor points. The "OR an equivalent demand for frequent or continuous muscular effort" part of the definitions appears to have less influence, although the "stamina" tasks (e.g., "drive cross-country in convoy" , which is mentally and physically demanding but requiring more stamina than strength), did appear further down the ranked list of physically demanding tasks. An alternative explanation may be that the PSE scale discriminates between strength and stamina better than does the PPE one, and that the stamina tasks really are less physically demanding. Assessment of metabolic cost in task performance would need to be conducted to determine if this alternative is true, but Ayoub, et al. (1982), using the PSE scale, Hogan, et al. (1979), using the PPE scale, and Sharp, et al. (1980), using actual metabolic cost, all found that about 90 percent of physically demanding tasks are so because weights lifted and carried. Also, the PPE scale is linked to metabolic cost via its development and the example tasks on the scale. This is an area that could be subject to more research. In practice, there is little problem in selecting "weight" tasks from the tasks identified by these job analysis procedures, so either scale would have served the purpose.

A further factor to emerge from the task analysis interviews is that those interviewed clearly preferred the PSE scale. Possibly this was because the PSE scale makes it clear, by using specific weights and heights, what the rating frame of reference should be; whereas the PPE scale is less clearly defined for these raters. This would suggest more face and content validity for the PSE scale than for the PPE one.

Implications for Future Job Analyses

This research suggests that the job analysis to identify important physically demanding tasks can be done satisfactorily using either pairs of scales. Scale reliabilities are good, there are high rank order correlations between the two physical demands scales and the two importance ones, and the task analysis failed to identify any significant physically demanding tasks not identified by the job analysis procedures.

Although these job analysis methods would be practical regardless of the task analysis methods used (i.e., metabolic cost, strength, or multiple physical abilities), job selection criteria that focus on strength could be justified on the basis of research (already cited) showing the dominance of strength requirements in performing physically demanding tasks. Use of such criteria offers economy of resources needed in selection testing and in conducting job analysis. Measurement of strength is far simpler than the measurement of stamina, and job analysis could focus on weights lifted and carried. If strength oriented task analysis procedures are selected (and thus concentrate on weights lifted and carried), then the PSE scale appears to offer advantages over the PPE one. It focuses rater attention on those characteristics by virtue of its scale point definitions; raters felt more comfortable using the PSE scale; and there were fewer tasks in the group selected for task analysis because the PSE scale tended to exclude "stamina" type tasks.

Regardless of which set of scales are selected for use, this research offers some very significant (and cost reducing) implications for the design of the job analysis. Reliability statistics from this research show very good results not only for large samples, but also for quite small ones. Satisfactory reliability coefficients were obtained for all scales from a sample size of k=50 raters. Since two to three hours of work time are saved for each worker not required to complete a survey questionnaire booklet, this can mean many man-years of work saved by applying these findings.

The high correlation between the panel members' ratings and those of the main survey on the PSE/CIP-9 scales, and the high reliability results for the panel ratings, suggest that the job analysis may be undertaken by using the experienced panel to develop a task inventory and to rate the tasks using the selected scales. (Although the PPE/CIP-7 scales were not tested in this way, their high correlations with PSE/CIP-9 suggest they would provide acceptable results.) If the interrater reliability results from the panel members is acceptable, these could be used to identify tasks for task analysis, thus saving the resources required for the survey. If the panel ratings do not demonstrate acceptable reliability, then a survey of approximately 50 experienced workers could be conducted. An alternative to the survey may be to vary the criteria used for selecting tasks for task analysis. One way to do this would be to accept the ratings for those tasks that show high rater agreement, and select more apparantly high demand (but also high rater variance) tasks for task analysis. Put another way, the job analysis is to ensure a complete coverage of the employment via the task inventory. The task ratings are to target the use of labour, skill and equipment intensive resources, and where there is more variance in the task ratings than is desirable, some compensation can be made by subjecting more tasks to task analysis. It is estimated that the successful application of the procedures described in this paragraph would save about ten man-years of Army project staff work if they were to be applied to the setting of physical demands criteria for Army jobs. Many man-years of worker time would also be saved.

References

Archer, W. B., & Fruchter, D. A. (1963). The construction, review, and administration of Air Force job inventories (PRL-TDR-62-21, AD-426 755). Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division.

Astrand, P-O., & Rhyming, I. (1954). A nomogram for calculation of aerobic capacity from pulse rate during sub-maximal work. Journal of Applied Physiology, 7, 218-221.

Astrand, P-O., & Rodahl, K. (1977). Textbook of Work Physiology. (2nd ed.) New York: McGraw-Hill.

Ayoub, M. M., Denardo, J. D., Smith, J. L., Bethea, N. J., Lambert, B. K., Alley, L. R., & Duran, B. S. (1982). Establishing physical criteria for assigning personnel to Air Force jobs (Final Report). Lubbock, TX: Texas Tech University, Institute for Ergonomics Research.

Campion, M. A. (1983). Personnel selection for physically demanding jobs: Review and recommendations. Personnel Psychology, 36, 527-559.

Chaffin, D. B. (1975). Ergonomics guide for the assessment of human static strength. American Industrial Hygiene Association Journal, 36, 505-511.

Christal, R. E. (1974). The United States Air Force occupational research project (AFHRL-TR-73-75). Lackland AFB., TX: Occupational Research Division.

Collyer, R. S., & James, R. F. (1985). Report on pilot research to identify and test a methodology for the setting of physical demands criteria for Army employments (Project Team Report). Canberra, ACT: Department of Defence, Directorate of Personnel Plans - Army (Limited distribution).

Consolazio, C.F. (1971). Energy expenditure studies in military populations using Kofranyi-Michaelis respirators. The American Journal of Clinical Nutrition, 24, 1431-1437.

Fine, S., & Wiley, W. (1971). An introduction to Functional Job Analysis (Methods for Manpower Analysis #4). Kalamazoo, MI: Upjohn Institute for Employment Research.

Fleishman, E. A., & Hogan, J. C. (1978). A taxonomic method for assessing the physical requirements of jobs: The Physical Abilities Analysis approach (Technical Report). Washington, DC: Advanced Research Resources Organization.

Goody, K. (1976). Comprehensive Occupational Data Analysis Programs (CODAP): Use of REXALL to identify divergent raters (AFHRL-TR-76-82). Lackland AFB., TX: Occupational and Manpower Research Division.

Gott, S. P., & Alley, W. E. (1980). Physical demands for Air Force occupations: A task analysis approach. Proceedings of the 22nd Annual Conference of the Military Testing Association. Toronto, Canada: Canadian Forces Personnel Applied Research Unit.

Healy, W. Occupational Analysis Report: Vehicle Mechanic Trades ECN 016, 223, 229 (Survey Report). (1984). Canberra, ACT: Department of Defence, Directorate of Personnel Plans -- Army, Military Employments Research and Information Team (MERIT) (Limited distribution).

Hogan, J. C., Ogden, G. D., Gebhardt, D. L., & Fleishman, E. A. (1979). Methods for evaluating the Physical and effort requirements of Navy tasks: Metabolic performance, and physical ability correlates of perceived effort (Contract No. N0014-78-C-0430). Washington, DC: Advanced Research Resources Organization.

Jansen, H. P. (1982). Identification of rating policies in Training Emphasis task factor data. Proceedings of the 24th Annual Conference of the Military Testing Association. San Antonio, TX: US Air Force Human Resources Laboratory and the US Air Force Occupational Measurement Center.

Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. Boston, MA: Houghton-Mifflin.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1969). Position Analysis Questionnaire (Purdue Research Foundation Contract No. 4497). West Lafayette, IN: Purdue Research Foundation.

Military Employments Research and Information Team (MERIT). (1982) Standing Operating Procedures, Canberra, ACT: Directorate of Personnel Plans -- Army.

Myers, D. C., Gebhardt, D. L., & Fleishman, E. A. (1980a). Development of physical performance standards for Army jobs: The job analysis methodology (USARIBSS-TR-446). Alexandria, VA: United States Army Research Institute for the Behavioural and Social Sciences.

Myers, D. C., Gebhardt, D. L., & Fleishman, E. A. (1980b). Physical performance standards for Army jobs: Criterion task manual (ARI-80-50). Alexandria, VA: United States Research Institute for the Behavioural and Social Sciences.

Sharp, D. S., Wright, J. E., Vogel, J. A., Patten, J. F., Daniels, W. L., Knapik, J., & Kowal, D. M. (1980). Screening for physical capacity in the US Army: An analysis of measures predictive of strength and stamina (USARIEM-T-8/80). Natick, MA: United States Army Research Institute of Environmental Medicine.

Appendix 1-1

Physical Strength and Endurance

Scale Point	Description of Effort
0 -	No Significant Physical Demand -- Corresponding requirement would include periodic lifting of 9 lbs or less -- includes most administrative and clerical tasks.
1 -	Extremely Light -- Corresponding requirement would include periodic lifting of 10 - 19 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
2 -	Very Light -- Corresponding requirement would include periodic lifting 20 - 29 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
3 -	Light -- Corresponding requirement would include periodic lifting of 30 - 39 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
4 -	Light to Moderate -- Corresponding requirement would include periodic lifting of 40 - 49 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
5 -	Moderate -- Corresponding requirement would include periodic lifting of 50 - 59 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
6 -	Moderate to Heavy -- Corresponding requirement would include periodic lifting of 60 - 69 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
7 -	Heavy -- Corresponding requirement would include periodic lifting of 70 - 79 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
8 -	Very Heavy -- Corresponding requirement would include periodic lifting of 80 - 89 lbs to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
9 -	Extremely Heavy -- Corresponding requirement would include periodic lifting of 90 lbs or more to a height of 5 ft OR an equivalent demand for frequent or continuous muscular effort.
X -	No knowledge of task requirement.

Appendix 1-2

PHYSICAL STRENGTH AND ENDURANCE

This scale is a measure of physical strength and endurance. Physical strength and endurance are defined as involving significant use of the "large' muscle groups in the arms, back or legs. These would include requirements for lifting, lowering or carrying heavy or cumbersome objects, pushing or pulling, torquing or any other demand for frequent or continuous exertion or muscular effort.

Remember, make your ratings on the basis of:

a. The most demanding aspects of each trade.
b. The level of demand placed on a single individual performing the task, and
c. The level of demand required by the complete task from start to finish.

1 - NONE OR EXTREMELY LIGHT PHYSICAL DEMAND - Corresponding requirement would include periodic lifting of 5 kilos (11 lbs) or less - includes most clerical and administrative tasks.

2 - VERY LIGHT - Corresponding requirement would include periodic lifting of 5 - 10 kilos (11 - 22 lbs) to a height of 1.5 metres OR an equivalent demand for frequent or continuous muscular effort.

3 - LIGHT - Corresponding requirement would include periodic lifting of 10 - 15 kilos (22 - 33 lbs) to a height of 1.5 metres OR and equivalent demand for frequent or continuous muscular effort.

4 - LIGHT TO MODERATE - Corresponding requirement would include periodic lifting of 15 - 20 kilos (33 -44 lbs) to a height of 1.5 metres OR an equivalent demand for frequent or continuous muscular effort.

5 - MODERATE - Corresponding requirement would include periodic lifting of 20 - 25 kilos (44 - 55 lbs) to a height of 1.5 metres OR an equivalent demand for frequent or continuous muscular effort.

6 - MODERATE TO HEAVY - Corresponding requirement would include periodic lifting of 25 - 30 kilos (55 - 66 lbs) to a height of 1.5 metres OR an equivalent demand for frequent or continuous muscular effort.

7 - HEAVY - Corresponding requirement would include periodic lifting of 30 - 35 kilos (66 - 77 lbs) to a height of 1.5 metres OR an equivalent demand for frequent or continuous muscular effort.

8 - VERY HEAVY - Corresponding requirement would include periodic lifting of 35 - 40 kilos (77 - 88 lbs) to a height of 1.5 metres OR an equivalent demand for frequent or continuous muscular effort.

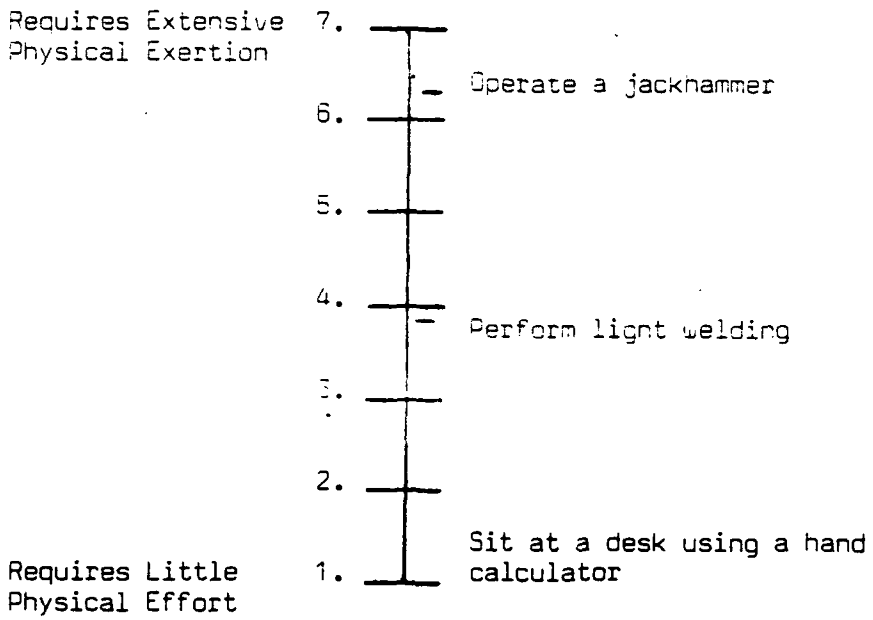
9 - EXTREMELY HEAVY - Corresponding requirement would include periodic lifting of more than 40 kilos (88 lbs) to a height of 1.5 metres OR an equivalent demand for frequent or continuous muscular effort.

X - No knowledge of task requirement.

Appendix 2

PERCEIVED PHYSICAL EFFORT

This is the degree of physical exertion experienced in performing a single task or a series of tasks.



Using the 7-point scale, please rate how much Effort it takes to perform each task.

Appendix 3

CONSEQUENCES OF INADEQUATE PERFORMANCE

This scale is a measure of the seriousness of probable consequences of inadequate performance of an activity. It is defined in terms of probable injury or death, failure to accomplish a critical mission, wasted supplies, damaged equipment, and wasted man-hours of work.

1 - EXTREMELY LOW CONSEQUENCES (neglible effect on people, equipment, mission)

2 - LOW CONSEQUENCES

3 - WELL BELOW AVERAGE CONSEQUENCES

4 - SOMEWHAT BELOW AVERAGE CONSEQUENCES

5 - AVERAGE CONSEQUENCES

6 - SOMEWHAT ABOVE AVERAGE CONSEQUENCES

7 - WELL ABOVE AVERAGE CONSEQUENCES

8 - HIGH CONSEQUENCES

9 - EXTREMELY HIGH CONSEQUENCES (may result in injury, death, serious damage to equipment, or failure to accomplish a critical mission).

Appendix 4

CONSEQUENCES OF INADEQUATE PERFORMANCE

This scale is a measure of the seriousness of probable consequences of inadequate performance of a job. It is defined in terms possible injury or death, wasted supplies, damaged equipment, and wasted man-hours of work. The work is to be rated on a scale from 1 (Least Serious Consequences of Inadequate Performance) to 7 (Most Serious Consequences of Inadequate Performance) with intermediate levels defined as follows:

What will happen if the job is inadequately performed?

7. Most serious consequences (eg check parachute rigging prior to personnel drop).
6.
5.
4. Moderately serious consequences (e.g., prepare ammunition for destruction).
3.
2.
1. Least serious consequences (eg fold hospital linen).

Using the 7-point scale, please rate what will happen if the task is inadequately performed.

END

1-87

DTIC