DEFENCE PHYSICAL EMPLOYMENT STANDARDS PROJECT
Infantry and Airfield Defence Guards

REPORT 10

RELIABILITY OF POTENTIAL PHYSICAL EMPLOYMENT TESTS: INFANTRY AND ADG

J T Harvey
W R Payne
W L Knez
D J Ham

May 2006

CONTRACT C538679

CONDUCT OF A PHYSICAL EMPLOYMENT STANDARDS STUDY FOR THE AUSTRALIAN DEFENCE FORCE

CONTACT
Professor Warren Payne
Project Manager
Defence Physical Employment Standards Project
School of Human Movement and Sport Sciences
University of Ballarat
PO Box 663 Ballarat Victoria 3353
Phone: 03 5327 9693
Fax: 03 5327 9060
Email: w.payne@ballarat.edu.au

DEFENCE PHYSICAL EMPLOYMENT STANDARDS PROJECT
Infantry and Airfield Defence Guards

COMPLETED AND PLANNED REPORTS

No.	Short Title[1]	Date[1]	Type
	Completed Reports		
1	Selection of Key Trade Tasks for Detailed Observation	Mar 04	Minor
2	Selection of Potential Endurance Tests & Anthropometric Measures	Sep 04	Minor
3	Review of Injury Data: Infantry and ADG	Feb 05	Minor
4	Trade Tasks Movement Analysis: Infantry and ADG	Apr 05	Minor
6	The Effect of Physically Demanding Infantry and ADG Trade Tasks on Cognitive Performance: a Pilot Observational Study	Apr 05	Minor
8	Selection of Criterion Trade Tasks: Infantry and ADG	Mar 05	Minor
10	Reliability of Potential Physical Employment Tests: Infantry and ADG	May 06	Minor
	Planned Reports		
5	Trade Tasks Risk Analysis and Risk Mitigation: Infantry and ADG	Jun 06	Minor
7	Retrospective Survey of Injuries: Infantry and ADG	Jul 06	Minor
9	Trade Task Analysis: Infantry and ADG	Jul 06	Major
11	Normative Physical Performance Data: Infantry and ADG	Jun 06	Major
12	Physical Performance Tests and Standards: Infantry and ADG	Jul 06	Major
13	Capacity of Women to Improve Physical Performance: a Review	Jul 06	Minor

[1] In the case of planned reports, both the titles and the dates of publication are provisional.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the assistance and support of Mr John Mathieson and Ms Michelle Dean, Defence Physical Employment Standards Project Office; MAJ Brett de Masson, Army Personnel; LT COL David McKerral, Army Trade Management; WO1 Peter Bradley, HQ Combined Arms Training Centre; WO1 Glenn Moorby, HQ training Command - Army; WOFF Nick Bandy, Combat Support Group; the officers of 1 RAR, 2 RAR, 3 RAR, 6 RAR, 1 JSU, 2 HSB, 7 CSSB and SOI; and the officers of AFDW and RAAFSFS. The authors would also like to express their appreciation to the soldiers and airmen who volunteered to participate in this project.

The authors also wish to acknowledge the contributions made by the following members of the DPESP Peer Review Panel, who critically reviewed a draft of this report: Dr John Brotherhood, Dr John Culvenor, Assoc Prof Leonie Otago, Ms Deb Pascoe, Dr Mark Rayson, Assoc Prof Steve Selig, Dr Bob Stacy, Ms Judy Swan, Ms Rebecca Tanner and Dr Chris Turville; and by Ms Jill Boatman to the production of this report.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

THE AUTHORS

Dr Jack Harvey is a Senior Research Fellow in the School of Information Technology and Mathematical Sciences at the University of Ballarat. He is a mathematical statistician with over 20 years experience in applied research in many contexts including human movement science, health sciences, occupational health & safety, and social and behavioural sciences. In the Defence Physical Employment Standards Project (DPESP), Dr Harvey has professional roles in research design, data management and statistical analysis, and is also Technical Manager of the project.

Professor Warren Payne is the Professor of Human Movement Science in the School of Human Movement and Sport Sciences at the University of Ballarat. He is an exercise physiologist with over 20 years of research and consulting experience. This experience has included working with a variety of groups and individuals from a range of backgrounds including elite athletes (rowing, cycling, badminton and swimming) and workers involved in heavy manual trades (sheep shearers, fire fighters, aircraft baggage handlers and plasterers). Professor Payne is the Project Manager and Research Leader of the DPESP.

Dr Wade Knez is a Research Fellow in the School of Human Movement and Sport Sciences at the University of Ballarat. He is an exercise physiologist and also has qualifications in psychology. He has eight years of experience working with athletes at all levels from recreational to elite professional. His specific area of expertise is field based studies. In the DPESP, Dr Knez's roles include leadership of field testing teams, movement analysis and cognitive analysis.

Daniel Ham is a Senior Research Assistant in the School of Human Movement and Sport Sciences at the University of Ballarat. He is an exercise physiologist and specialises in endurance performance and fatigue. He has a Bachelor of Human Movement degree and is currently completing Honours in exercise physiology. In the DPESP, Mr Ham's role is in the collection and analysis of field data.

EXECUTIVE SUMMARY

Background

Military operational tasks are physically demanding and incur the risk of injury. In order to address the issues and costs associated with the high injury rates and focus on ways to reduce the risk of injury to Australian Defence Force (ADF) personnel, the ADF Chiefs of Service Committee (COSC) has endorsed a number of injury prevention strategies aimed at examining, analysing and evaluating injury-related risks and hazards within the ADF. In line with those strategies, COSC has affirmed that ADF employment policy is to be competency based and agreed that physical employment standards should be developed for combat arms trades. The purpose of the Defence Physical Employment Standards Project (DPESP) is to develop these performance-based competency standards.

The ADF has employed the services of the University of Ballarat (UB) to undertake the DPESP. This involves reviewing combat arms trade tasks (CATTs), establishing a set of criterion CATTs, developing a battery of simulation and predictive tests based on the criterion CATTs to be used to assess the physical competency of ADF combat personnel, and making recommendations for associated physical employment standards.

In the initial phase, the study is focused on one Army corps - Infantry, and one Air Force mustering - Airfield Defence Guards (ADG).

Aim

The aim of this work package was:

- to undertake development, including pilot testing, of the proposed new physical employment tests (PETs) recommended by the Criterion Tasks Workshop, as outlined in DPESP Report 8;
- to establish the reliability of the developed tests (i.e. to establish equitable and defensible tests which produce consistent results when the same individuals are tested on different occasions and by different testers); and
- to evaluate test reliability in each of four cohorts: trained infantry, infantry initial employment training (IET), trained ADG and female Army personnel.

Tests to be Evaluated

The set of tests recommended by the Criterion Tasks Workshop fell into two categories. The first category consisted of standard fitness battery tests in common use, for which both protocols and reliability are well established. The second category consisted of new tests requiring development, refinement, and for which reliability needed to be established.

The standard fitness battery tests included:

- Heaves
- Jump and Reach
- Sit-ups
- Grip Strength
- Unloaded Shuttle Run

The reliability of these tests has been established by previous research. No test development or reliability testing was undertaken for these tests in this study.

The new tests included simulations of criterion tasks and predictive tests designed to predict performance on criterion tasks.

The tasks for which simulation tests were to be developed included tasks common to both Infantry and ADG, and tasks specific to either Infantry or ADG. These were:

Tasks common to Infantry and ADG
- Load and Unload UNIMOG
- Jerry Can Carry
- 1.82m Wall Climb
- Tunnel Crawl
- Urban Rushing
- Section Attack

Tasks for ADG only
- Sustained Patrol (5km)
- Pursuit (2.4km)

The predictive tests to be developed were:

Tests common to Infantry and ADG
- 1.82m Wall Climb from a standing start (to be developed in conjunction with the simulation test which is based on a running start).
- Loaded Incremental Velocity Run

These two tests were considered to have generic potential to predict performance on a range of criterion tasks.

Test for Infantry only
- Forced March (10km). This was included for the specific purpose of predicting performance on the Forced March (20km) criterion task.

It was also recommended by the Criterion Tasks Workshop that the 20 km and 10 km Forced Marches for Infantry should be performed immediately before the Section Attack simulation test, and should be considered as pre-fatiguing endurance activities for the Section Attack test, rather than as tasks/tests in their own right.

Following further discussions subsequent to the Criterion Tasks Workshop, it was also agreed with ADG that the Pursuit should be performed immediately before the Section Attack simulation test, and should be considered as a pre-fatiguing endurance activity for the Section Attack test, rather than as a task/test in its own right.

Methodology

Test development and reliability assessment were carried out under protocols approved by Defence and University of Ballarat Human Research Ethics Committees. Test protocols were developed and refined in consultation with key Infantry and ADG informants. Pilot testing took place during October 2004 and involved 39 volunteers: 20 soldiers from 3 RAR and 19 airmen from AFDW. Reliability testing took place with volunteers from AFDW and 6 RAR, and volunteer female Army personnel from 2 HSB, 1 JSU and 7 CSSB (Land Command Brisbane Military Area), during the period 15 Nov - 3 Dec 2004, and with volunteer soldiers in Initial Employment Training (IET) at the School of Infantry at Singleton during the period 7-11 Mar 2005. The sample of 65 was made up of: 20 Infantry soldiers, 19 ADG airmen, 10 female Army personnel and 16 Infantry IETs.

Assessing the reliability of a physical employment test (PET) involves testing the same people on two or more occasions over a period of days, to examine the level of agreement (or reproducibility) of test results from different occasions. In order to achieve this, whilst ensuring that there was adequate recovery time after all tests, and in particular after the more demanding tests, the sample at each location was divided into two groups, with half of the tests being allocated to each group. The final sample sizes for each test ranged from 21 to 30.

The reliability of each test was assessed on the basis of standard statistical measures including intraclass correlation (ICC), bias, limits of agreement (LOA) and Cohen's kappa statistic.

In physical tests which are administered and scored by an individual tester or a team of testers it is of interest to establish both intra-tester reliability (the same tester on two or more occasions)[1] and also inter-tester reliability (different testers on each occasion). The latter is particularly important for establishing generalisability, which is crucial for operational field-based testing. It can be argued that intra-tester reliability is a necessary pre-condition for inter-tester reliability, and hence that establishing the latter implicitly also establishes the former. In this study, there were insufficient physical training instructors (PTIs) available at each location for inter-tester reliability to be assessed in a pure and unconfounded manner. In most cases, the functions of individual members of the testing team were rotated between trials to the degree that this was feasible. Hence the reported reliabilities represent an amalgam of inter-tester and intra-tester reliability.

Results

The results of reliability testing are summarised in the table on the following page.

Two things should be noted:

- During the testing program, it was decided for reasons of both safety and reliability, to replace the Pack Lift and Place test by a Box Lift and Place test, closely modeled on a test used in the UK.
- In tests on which performance includes a skill or confidence component, reliability is improved by making adequate provision for familiarisation. On the advice of Defence informants, it was assumed that all participants would be familiar with the commonly performed CATTs on which the simulation tests were based. Whilst this was the case with the participants in the pilot testing, it became apparent during reliability testing that in most cases participants would have benefited from more extensive opportunity for familiarisation than was possible within the time constraints of the DPESP reliability field trials. Tests for which some cohorts particularly require familiarisation are indicated.

Conclusions

Considering the practical constraints on the conditions of testing discussed below, the levels of reliability of most tests are regarded as either acceptable or provisionally acceptable.

The best reliabilities, as indicated by low LOAs, were generally obtained with IETs at SOI, suggesting that this cohort in particular tended to put in very consistent efforts. In the CATT-based tests, IETs tended to improve as a group from trial 1 to trial 2. This did not occur in the generic (ie. non CATT-based) Loaded Incremental Velocity Run test. This supports the notion that the general group improvement effect observed in the CATT-based tests had a substantial learning component. This in turn indicates the need for thorough familiarisation with CATT-based tests in the normative testing phase. In the case of exhaustive tests, this requires a separate familiarisation session at least 48 hours prior to testing. This should be taken into consideration when planning the schedule of normative data collection with IETs from both Infantry and AFDW.

Some of the results of the first round of reliability testing at AFDW appeared to be affected by the requirement for ADGs to become familiar with tasks that are performed less regularly by ADG than Infantry. Results from 6 RAR were affected by problems with attitude and compliance with protocols on the part of PTIs, and by the lack of motivation and poor physical condition of the soldiers, many of whom were carrying undisclosed injuries from previous Infantry training activities.

Considering these limitations, it is considered that the results of this study indicate that the reliability of most of the PETs is acceptable or provisionally acceptable. However, the moderate reliability established for a number of these tests indicates that there is the potential for a substantial degree of variation to occur from occasion to occasion in operational use. For this reason it is imperative that adequate familiarisation occurs prior to testing, and that there is adequate opportunity for retesting if a test is failed. It is anticipated that the former will occur since the tests will be used as a basis for training, and that the latter will occur as a matter of course within the competency testing paradigm.

[1] Intra-tester reliability is a special case of test-retest reliability.

Summary of Results of Reliability Testing

Test Name	Infantry soldiers	AFDW airmen	Infantry IETs	Army females	Overall Assessment
Simulation tests					
Tests common to Infantry and ADG					
Pack Lift and Place	×	×	NT	NT	U
Box Lift and Place	NT	NT	✓	NT	PA
Jerry Can Carry	✓	✓	✓	×	A
1.82m Wall Climb (running start) Categorical test: pass/fail	+	+	✓	✓	A
1.82m Wall Climb (running start) Quantitative test: duration	✓	✓	✓ Famil. req.	?	A
Leopard Crawl	✓	✓	✓ Famil. req.	✓ Famil. req.	A
Urban Rushing	?	?	✓ Famil. req	× Famil. req	PA
Section Attack	?	?	✓ Famil. req.	NT[1]	PA
Tasks/Tests for ADG only					
Sustained Patrol (5km)	NA	+	NA	NA	ND
Pursuit (2.4km)	NA	+	NA	NA	ND
Predictive tests					
Tests common to Infantry and ADG					
1.82m Wall Climb (standing start) Categorical test: pass/fail	+	+	✓	–	PA
1.82m Wall Climb (standing start) Quantitative test: duration	×	×	×	–	U
Loaded Incremental Velocity Run	✓	✓	✓	NT	A
Test for Infantry only					
Forced March (10km)	+	NA	+	NT[1]	ND

[1] Females undertook these tests on only one occasion, so no reliability analysis was possible.

Key:

✓ Reliability satisfactory	+ All participants passed test	A Reliability acceptable
× Reliability not satisfactory	– All participants failed test	U Reliability unacceptable
? Reliability marginal or uncertain	NA Not applicable	PA Reliability provisionally acceptable
	NT Not tested	ND No discrimination - no reliability assessment

Because of the small number of females recruited for reliability testing, and the fact that not all those recruited were physically capable of safely undertaking the more physically demanding tasks, it was not possible to complete a full program of testing on women in the limited time available. Of the five tests for which reliability assessments were possible for this cohort, one (Leopard Crawl) was found to be reliable (but requiring thorough familiarisation), and one (Wall Climb with Running Start) was found to be marginally reliable. No females were able to complete a Wall Climb with Standing Start. Reliability was not established for two tests (Jerry Can Carry and Urban Rushing). The data for females exhibited both changes from trial to trial in the performance level of the group as a whole, and large trial to trial variation in individuals. The former is probably indicative of learning effects; the latter may be related to both unfamiliarity with the tasks and how to approach them, and to issues with poorly fitting equipment and apparel, particularly ballistic vests and helmets.

Notwithstanding the previous paragraph, it should be noted that in the three tests completed by females, the level of performance of females was significantly lower than that of all the male cohorts. In two of these tests (jerry can carry and leopard crawl) there was no overlap at all in the male and female scores - the best individual female performance was worse than the worst individual male performance. In the other test (running wall climb), the best female performances were comparable with the worst male performances (in the SOI cohort). In general, in all five tests undertaken by females, the differences between the data for males and females were so marked that the existence of underlying differences on these tests can be confidently asserted even though the female performances were generally measured with lower reliability than the male performances. However it is not possible to draw any conclusion as to the degree to which these results reflect inherent gender differences in physical capacity. Females in this study were less engaged in physical training on a day-to-day basis than were participants in the three male cohorts, and had had less exposure to the combat trade tasks on which most of the tests were based. As a result, gender differences are confounded with the effects of training and familiarisation. However, it should also be noted that there were two levels of selectiveness operating with the female cohort. Firstly, whilst all participants were volunteers, the element of self selection was much stronger with females, for whom the testing program was unrelated to their current employment category and peer group; and secondly, there was a two-stage fitness hurdle controlling entry into the test program. These mechanisms would presumably have biased the female sample towards better than average female performance.

Recommendations

The reliability of the following tests (common to Infantry and ADG) is assessed as acceptable for adoption as PETs. It is recommended that data on these tests be collected in the Normative Data Collection phase of the DPES Project:

- Jerry Can Carry
- 1.82m Wall Climb from Running Start (categorical and quantitative: pass/fail and duration)
- 1.82m Wall Climb from Standing Start (categorical test: pass/fail)
- Leopard Crawl
- Loaded Incremental Velocity Run

Whilst the results for the following tests at SOI were encouraging, overall the reliability of these tests could only be assessed as provisionally acceptable. It is recommended that data on these tests be collected in the Normative Data Collection phase of the DPES Project. However, it is also recommended that if any of these tests is adopted by Defence, further reliability testing should be conducted with thoroughly familiarised and motivated participants in good physical condition, in order to confirm the provisional conclusions reached in this report.

Tests common to Infantry and ADG
- Box Lift and Place
- Urban Rushing

Test for Infantry only
- Section Attack after Forced March (10km)

Test for ADG only
- Section Attack after Pursuit (2.4km)

The reliability of the following ADG test could not be assessed because all participants passed the test on both occasions and hence the capacity of the test to discriminate was not established. It is recommended that this test not be further considered in the Normative Data Collection phase of the DPES Project.

- Sustained Patrol (5km)

The reliability of the following tests is assessed as unacceptable for adoption as PETs. It is recommended that these tests not be further considered.

- Pack Lift and Place
- Wall Climb from Standing Start (quantitative test: duration).

In order to optimise the reliability of the data to be collected from IE trainees and females in the forthcoming normative phase of the DPES Project, it is recommended that an extensive program of familiarisation with CATT-based tests should be undertaken by these cohorts prior to actual testing. This requirement should be taken into consideration when planning the schedules for normative data collection for IETs from both Infantry and AFDW, and for females.

It is recommended that steps are taken by Defence to ensure that female participants are equipped with well-fitting ballistic vests & helmets during both familiarisation and normative testing.

REFERENCE DOCUMENTS

A. Commonwealth of Australia. (2002). *Request for Tender for Conduct of a Physical Employment Standards Study for the Australian Defence Force, Part One: Draft Statement of Work.* Canberra.

B. Commonwealth of Australia. (2003). *Contract C538679 Conduct of a Physical Employment Standards Study for the Australian Defence Force*. Canberra.

C. Stacy, R.J., Payne, W.R. and Harvey, J.T. (2004). *Defence Physical Employment Standards Project, Infantry and Airfield Defence Guards; Report 1: Selection of Key Trade Tasks for Detailed Observation*. Canberra: Department of Defence, Defence Personnel Executive.

D. Payne, W.R, Brotherhood, J.R., Harvey, J.T., Knez, W.L., Kay, B., and Selig, S.E. (2004). *Defence Physical Employment Standards Project, Infantry and Airfield Defence Guards, Report 2: Selection of Potential Endurance Tests & Kinanthropometric Measures.* Canberra: Department of Defence, Defence Personnel Executive.

E. Payne, W.R., Knez, W.L., Harvey, J.T., Sinclair, W.H., Elias, G.P. and Ham, D.J. (2005). *Defence Physical Employment Standards Project, Infantry and Airfield Defence Guards; Report 4: Trade Tasks Analysis: Infantry and ADG*. Canberra: Department of Defence, Defence Personnel Executive.

F. Payne, W.R., Harvey, J.T., Knez, W.L, and Ham D.J. (2005). *Defence Physical Employment Standards Project, Infantry and Airfield Defence Guards, Report 8: Selection of Criterion Trade Tasks: Infantry and ADG*. Canberra: Department of Defence, Defence Personnel Executive.

ABBREVIATIONS AND ACRONYMS

ADF	Australian Defence Force
ADG	Airfield Defence Guard
ADHREC	Australian Defence Human Research Ethics Committee
AFDW	Airfield Defence Wing
CATT	Combat Arms Trade Task
COSC	Chiefs of Service Committee
DPESP	Defence Physical Employment Standards Project
CV	Coefficient of Variation
ICC	Intraclass Correlation
IET	Initial Employment Training
LOA	Limits of Agreement
PES	Physical Employment Standards
PET	Physical Employment Test
PTI	Physical Training Instructor
SEM	Standard Error of Measurement
SOI	School of Infantry
UBHREC	University of Ballarat Human Research Ethics Committee
1 JSU	1st Joint Support Unit
2 HSB	2nd Health Support Battalion
2 RAR	2nd Battalion, Royal Australian Regiment
3 RAR	3rd Battalion, Royal Australian Regiment
6 RAR	6th Battalion, Royal Australian Regiment
7 CSSB	7th Combat Service Support Battalion

1 INTRODUCTION

1.1 Background

1.1.1 Military operational tasks are physically demanding and incur the risk of injury. In order to address the issues and costs associated with the high injury rates and focus on ways to reduce the risk of injury to Australian Defence Force (ADF) personnel, the ADF Chiefs of Service Committee (COSC) has endorsed a number of injury prevention strategies aimed at examining, analysing and evaluating injury-related risks and hazards within the ADF. In line with those strategies, COSC has affirmed that ADF employment policy is to be competency based and agreed that physical employment standards should be developed for combat arms trades. The purpose of the Defence Physical Employment Standards Project (DPESP) is to develop these performance-based competency standards.

1.1.2 The ADF has employed the services of the University of Ballarat (UB) to undertake the DPESP. This involves reviewing combat arms trade tasks (CATTs), establishing a set of criterion CATTs, developing a battery of simulation and predictive tests based on the criterion CATTs to be used to assess the physical competency of ADF combat personnel, and making recommendations for associated physical employment standards (See Reference Documents A and B).

1.1.3 In the initial phase, the study is focused on one Army corps - Infantry, and one Air Force mustering - Airfield Defence Guards (ADG).

1.1.4 The stages of the DPESP study are:

a. identification and observation of CATTs;
b. analysis of physical demands and cognitive effects of CATTs;
c. identification and analysis of injury risks of CATTs;
d. identification of criterion CATTs on which to base tests of physical performance;
e. development of a set of potential physical employment tests (PETs), and establishment of their reliability and validity;
f. collection of normative data on the PETs;
g. selection of the final battery of PETs and determination of minimum performance standards on each.

1.2 Aims

1.2.1 This work package (WBS 1.4.2) relates to 1.1.4.e. The aims were:

a. to undertake development, including pilot testing, of the proposed new PETs recommended by the Criterion Tasks Workshop, as outlined in Report 8 (Reference Document F);
b. to establish the reliability of the developed tests (i.e. to establish equitable and defensible tests which produce consistent results when the same individuals are tested on different occasions and by different testers); and
c. to evaluate test reliability in each of four cohorts: trained infantry, infantry initial employment training (IET), trained ADG and female Army personnel.

1.3 Scope

1.3.1 The research design for reliability testing as outlined in the project methodology (see Reference Document B) specified 20 participants in each of four cohorts: trained infantry, infantry initial employment training (IET), trained ADG and female Army personnel; and a one week testing period for each cohort. One week with 10 soldiers at an unspecified location was also allocated for pilot testing. The target number of task-based tests to be examined within this framework was seven.

2 METHODOLOGY

2.1 Test Identification

2.1.1 The initial identification of potential tests was undertaken in the Criterion Tasks Workshop. A number of recommendations regarding test development were made by the Workshop and outlined in Report 8 of this series (Reference Document F). The set of tests recommended by the Criterion Tasks Workshop fell into two categories. The first category consisted of standard fitness battery tests in common use, for which both protocols and reliability are well established. The second category consisted of new tests requiring development, refinement, and for which reliability needed to be established.

2.1.2 The standard fitness battery tests included:

- Heaves
- Jump and Reach
- Sit-ups
- Grip Strength
- Unloaded Shuttle Run

The reliability of these tests has been established by previous research, as indicated in Table 1. No test development or reliability testing was undertaken for these tests in this study.

Table 1. Reliability of Standard Fitness Battery Tests

Test	Reference	Reliability Measures[1]			Sample	Notes
		Intraclass Correlation	Relative bias (% of mean)	Relative Limits of Agreement (% of mean)		
Shuttle run	Cooper et al. (2005)	-	1%	5%	21 males 21.8 ± 3.6 yrs	Reliability calculated for estimated VO_{2max}
Jump & reach	Young et al. (1997)	0.93	-	13%	17 males 18-25 yrs	LOA imputed from CV (see paragraph 2.5.5)
Situps	Sparling et al., (1997)	0.86	-	-	167 male & 38 female college students	25 reps/min protocol vs 20 reps/min (DPESP)
Grip strength	Mathiowetz (2002)	0.90-0.97	-	-	30 males 23-48 yrs	Inter-instrument reliability
Heaves	Engelman & Morrow (1991) Pate et al. (1993)	0.88 0.80	-	-	242 males grades 3-5 38 males 9-10 yrs	Very young cohorts

[1] The three measures of reliability are discussed in Section 2.5.

2.1.3 The new tests included simulations of criterion tasks and predictive tests designed to predict performance on criterion tasks. The tasks for which simulation tests were to be developed included tasks common to both Infantry and ADG, and tasks specific to either Infantry or ADG. These were:

Tasks common to Infantry and ADG

- Load and Unload UNIMOG
- Jerry Can Carry
- 1.82m Wall Climb
- Tunnel Crawl
- Urban Rushing
- Section Attack

Tasks for ADG only
- Sustained Patrol (5km)
- Pursuit (2.4km)

2.1.4 The predictive tests to be developed were:

Tests common to Infantry and ADG
- 1.82m Wall Climb from a standing start (to be developed in conjunction with the simulation test which is based on a running start).
- Loaded Incremental Velocity Run

These two tests were considered to have generic potential to predict performance on a range of criterion tasks.

Test for Infantry only
- Forced March (10km) This was included for the specific purpose of predicting performance on the Forced March (20km) criterion task

2.1.5 It was also recommended by the Criterion Tasks Workshop that the 20 km and 10 km Forced Marches for Infantry should be performed immediately before the Section Attack simulation test, and should be considered as pre-fatiguing endurance activities for the Section Attack test, rather than as tasks/tests in their own right.

2.1.6 Following further discussions subsequent to the Criterion Tasks Workshop, it was also agreed with ADG that the Pursuit should be performed immediately before the Section Attack simulation test, and should be considered as a pre-fatiguing endurance activity for the Section Attack test, rather than as a task/test in its own right.

2.2 Test Validity

2.2.1 For performance tests in the context of military training, the main validity requirements are for content validity and predictive validity (Boldovici et al., 2001). Content validity concerns how well the test represents the domain of performance included in the training. This kind of validity is established via well documented task analyses performed by subject-matter experts for the tasks included in the training (Boldovici et al., 2001; Goldstein et al.,1993). Predictive validity refers to the ability of test scores to relate to and forecast later measures of important behaviour. Measures obtained in subsequent training or exercise events provide criteria for partially validating Army performance tests. The ultimate criterion of validity for tactical tests is of course combat performance, and measures of this kind are virtually unobtainable.

2.2.2 In the DPESP project, content validity has been established through extensive consultation with subject matter experts and knowledge gained from the Task Observation and Analysis phase of the project (see Reference Documents C, D, E, and F).

2.2.3 Predictive validity of the predictive tests will be measured by empirically relating the performance on the predictive test to performance on the criterion task, using data to be gathered during the normative testing phase of the project.

2.2.4 A third aspect of validity is construct validity, which refers to the extent to which a test measures some underlying abstract construct, or in the physical testing context, some specific physiological trait or capacity. This has been addressed at a conceptual level by an analysis of muscle groups and actions associated with each test (see Reference Document F). It will also be possible to empirically determine construct validity by relating performance on the task-based tests to some of the laboratory-type tests. The capacity to do this is limited by the limited range of the battery of laboratory-type tests.

2.3 Test Development and Pilot Testing

2.3.1 Protocols for recruiting subjects for both developmental pilot testing and reliability testing, and protocols for the actual testing, were approved by the Australian Defence Human Research Ethics Committee (ADHREC) and the University of Ballarat Human Research Ethics Committee (UBHREC). The tests are listed in Table 2.

2.3.2 Preliminary pilot testing took place with personnel from 3 RAR on 11-15 Oct 2004 and AFDW on 26-27 Oct 2004. A total of 20 Infantry soldiers and 19 ADG airmen from nominated sections volunteered to participate after receiving a request to take part in the project from a member of the ADF who was not in the individual soldier's or airman's direct chain of command. Each volunteer was given an information sheet (Annex 5) and signed a consent form (Annex 6). Each soldier or airman also completed a confidential health status form (Annex 7). Soldiers or airmen who reported a pre-existing injury or illness that was likely to place them at an unacceptable degree of injury risk as a result of participation in the project were excluded from the project by the Chief Investigator. Participants at each location were divided into two groups, with each group performing half of the proposed tests. Each test was repeated and adjusted until it was considered by the researchers and the Defence subject matter experts that an appropriate standardised protocol had been established.

2.3.3 It was recognised that soldiers in IET at SOI and females were potentially at greater risk of injury in some of the most physically demanding endurance tasks than were experienced soldiers and airmen. Measures were taken to safeguard these groups wherever it was considered necessary.

2.3.4 In the case of soldiers in IET, with the approval of ADHREC and UBHREC, the total weight carried in the forced march was reduced from 45 kg to 25 kg. This load is equal to that normally carried by Infantry soldiers in IET and 5 kg heavier than that which is required to be carried over a slightly longer distance at a similar pace by recruit trainees (2005 Army Recruit Training exit standard: 20 kg for 12 km at 11min/km).

Table 2. Criterion Tasks, Simulation Tests and Predictive Tests

Task name	Test Name
	Simulation tests
Tasks/Tests common to Infantry and ADG	
Load and Unload UNIMOG	Pack Lift and Place
Jerry Can Carry	Jerry Can Carry
1.82m Wall Climb (running)	1.82m Wall Climb (running start)[1]
Tunnel Crawl	Leopard Crawl
Urban Rushing	Urban Rushing
Section Attack	Section Attack
Tasks/Tests for Infantry only	
Forced March (20km)	Forced March (20km)
Tasks/Tests for ADG only	
Sustained Patrol (5km)	Sustained Patrol (5km)
Pursuit (2.4km)	Pursuit (2.4km)
	Predictive tests
	Tests common to Infantry and ADG
	1.82m Wall Climb (standing start)[1]
	Loaded Incremental Velocity Run
	Test for Infantry only
	Forced March (10km)

[1] For the Wall Climb, both running and standing starts were incorporated in a single test protocol.

2.3.5 In the case of female soldiers, there is a tension between the requirement for the PES study to assess female capacity to undertake combat arms trade tasks, and the risk of injury.

2.3.6 In order to provide positive predictive value of female capacity to undertake combat arms trade tasks, females should be tested carrying the full load of 45 kg. Testing with a lighter load will provide negative predictive value (i.e. failure to pass the test will indicate inability to do the task) but not necessarily positive predictive value (since passing the test will not necessarily indicate ability to do the task).

2.3.7 However, from the perspective of reducing the risk of injury to females engaged in the study, a reduced load for female soldiers was considered more appropriate. With the approval of ADHREC and UBHREC, the load was set at 20 kg. This was based on two separate and independent considerations. Firstly, it matches the recruit exit standard for both male and female recruit trainees (2005 Army basic training exit standard: 20 kg for 12 km at 11min/km). Secondly, using data gathered in the task observation phase of the DPESP and applying the formula of Pandolf et al. (1977), the average male soldier carrying a load of 45 kg and walking at 6 km hr^{-1} is working at 68% of aerobic capacity (68% of 3.95 $l.min^{-1}$ or 2.69 $l.min^{-1}$). According to a recent UK MOD report, the average absolute aerobic capacity of female soldiers was approximately 67% of the average absolute aerobic capacity of male soldiers (i.e. 2.64 $l.min^{-1}$)(Women in the Armed Forces Steering Group, 2002). Applying the formula of Pandolf et al. (1977), a female soldier with this aerobic capacity and average weight (60 kg) carrying a load of 20 kg and walking at 6 km hr^{-1} is calculated to have a relative energy expenditure of 1.66 $l.min^{-1}$ or 63% of their capacity, which is similar to that of the average male soldier with a load of 45 kg (68%).

2.3.8 Performance on a multistage shuttle run was used to identify female soldiers with this level of aerobic capacity. It is calculated that a female soldier of age 21-30 and average weight (60 kg) who has achieved level 9-2 on the multistage shuttle run will have approximately 67% of the absolute aerobic capacity of the average male soldier, and hence when carrying a load of 20 kg will have a relative energy expenditure equal to that of the average male soldier with a load of 45 kg.

2.3.9 The relationship between injury risk and physical fitness is further supported by research conducted by Knapik et al. (1997) and the findings of an extensive review by Deuster et al. (1997), who have indicated that the most clearly identified predictor of injury risk in female military personnel when undertaking military training is cardiorespiratory fitness. For example, published data have demonstrated that female recruits with a two mile (3.2 km) run time greater than 19.07 min have a higher likelihood (risk ratio of 2.3) of incurring a musculoskeletal injury than those with a run time of less than 17.38 min (Knapik et al., 1997) and female soldiers completing a 1.6 km run in greater than the median time of 9.75 min have double the incidence of time-loss injuries compared to those who completed the run in less than 9.75 min (Jones et al., 1992).

2.3.10 Consequently, in order to ensure that female participants had a satisfactory level of fitness to safely undertake the testing program, they were divided into two groups. All female participants were required to have passed the BFA and to have undertaken a multistage shuttle run test within the previous six months. Those who had achieved level 9-2 participated in the endurance tests: 10 km Forced March, Section Attack and Urban Rushing. Those who had not achieved this level, but had satisfactorily completed the BFA within the previous six months, were tested on the Wall Climb, Leopard Crawl and Jerry Can Carry. For reasons of practicality, it was decided to administer the shuttle run test on the first of the five days of reliability testing. This precautionary approach was approved by ADHREC and UBHREC. It should be noted that any conclusions reached with regard to females will only apply to the sub-populations tested.

2.3.11 Even with these measures in place, the scarcity of female volunteers at both levels of fitness made it impossible to safely complete the whole reliability testing program within the one week scheduled. As a further risk reduction strategy, it was decided that females would not participate in the Pack Lift and Place or the Loaded Incremental Velocity Run. In the case of the Pack Lift and Place, problems had already been identified with the proposed test at AFDW and 6 RAR. It proved difficult to standardise the pack types, pack adjustments, pack contents, weight distribution and the manner of holding and lifting the packs, so it was subsequently replaced by the Box Lift and Place test. Among the more demanding tasks to be undertaken by the fitter females, reliability testing of the Loaded Incremental Velocity

Run, an extra predictive test, had lower priority than the three tests which were direct simulations of CATTs, and so it was omitted. It subsequently transpired that even this reduced testing schedule for the fitter females could not be completed – after four sessions in four days (Shuttle Run, Forced March and Section Assault, and two sessions of Urban Rushing) all participants were too sore and/or fatigued to repeat the Forced March and Section Assault on the fifth day.

2.3.12 Experience during pilot testing, and consultation with key Defence informants, led to a number of minor adjustments to the provisional specifications proposed at the Criterion Tasks Workshop for a number of the tests, as outlined in Report 8 (Reference Document F). Descriptions of the final form of each test are given in overview in the Test Procedures document (Annex 8). The full detailed protocols are in Annex 9.

2.3.13 Issues of safety and reliability emerged in pilot testing with the Pack Lift and Place Test. It proved difficult to standardise the pack types, pack adjustments, pack contents, weight distribution, the manner of holding and lifting the packs, and hence it was also difficult to standardise judgments about good form and stopping criteria. There were also safety issues associated with pack lifting and carrying techniques, and with the repetitive nature of the test as it was originally conceived. Modifications were made in an attempt to overcome these problems, but ultimately it became apparent during reliability testing that substantial problems remained, and it was decided to reject the test on both reliability and safety grounds, and instead adopt a UK one repetition maximum (1RM) Box Lift test. It is recognised that this test does not simulate repetitive lifting and carrying, but safety considerations required that a degree of content validity be sacrificed.

2.3.14 Notwithstanding the pilot testing, problems with some tests continued to emerge during reliability testing, related to equipment variations, and the injury status and motivation of participants. These matters are discussed further in Section 4.1.

2.4 Reliability Testing

2.4.1 Reliability testing took place with volunteers from AFDW and 6 RAR, and volunteer female Army personnel from 2 HSB, 1 JSU and 7 CSSB (Land Command Brisbane Military Area), during the period 15 Nov - 3 Dec 2004, and with volunteer soldiers in Initial Employment Training (IET) at the School of Infantry at Singleton during the period 7-11 Mar 2005.

2.4.2 The research design specified 20 participants in each of the four cohorts, with one week for reliability testing at each location. Because the number of tests to be examined at each location was greater than the seven targeted in the project methodology, and because a number of the tests were very physically demanding and required adequate time for recovery, it was not feasible for a particular participant to perform all tests on two or more occasions. Instead, the cohort of participants at each location was divided into two groups of nominal size 10, and the set of tests was allocated between the two groups. Thus the total sample size specified was 80, with a sample size of 40 for each test. The allocation of tests to groups and the order of testing was not randomized; rather, a mix of more physically demanding and less physically demanding tests was allocated to each group, and the schedule was structured to ensure that there was adequate recovery time after all tests, and in particular after the more demanding tests. The testing schedules for the four cohorts are attached as Annexes 1-4.

2.4.3 Due to practical constraints discussed below (Section 5.1), the reliability testing program was limited, both qualitatively because various confounding factors could not be controlled, and quantitatively since the target sample sizes were not achieved. The initial sample sizes achieved were: 20 Infantry soldiers, 19 ADG airmen, 10 female Army personnel and 16 Infantry IETs. As for the pilot testing, participants volunteered after receiving a request to take part in the project from a member of the ADF who was not in the individual soldier's or airman's direct chain of command. Each volunteer was given an information sheet (Annex 5) and a set of test procedures (Annex 8) and signed a consent form (Annex 6). Each soldier or airman also completed a confidential health status form (Annex 7). Soldiers or airmen who reported a pre-existing injury or illness that was likely to place them at an unacceptable degree of injury risk as a result of participation in the project were excluded

from the project by the Chief Investigator. Only one member of the nominated Infantry, Infantry IET and ADG units did not volunteer. However, there was a substantial drop out rate, particularly at 6 RAR, where it became apparent that a number of volunteers had undisclosed pre-existing injuries or medical conditions. Hence the effective sample sizes actually achieved for each test were lower than the target of 40, ranging from 21 to 30 (see Tables 6 and 7).

2.4.4 With regard to one of the newly developed simulation tests, the Section Assault, some preliminary data had also been collected in May 2004, during the task observation phase of the DPES Project. These data were collected from eight volunteers at 2 RAR under similar protocols to those referred to in paragraph 2.3.1, which had also been approved by both ADHREC and UBHREC. However, helmets and ballistic vests were not worn, and there was no pre-fatiguing march, although the trial took place late in the day after performance of a number of CATTs. These data were included in the reliability analysis.

2.4.5 A summary of the testing program is shown in Table 3. All relevant tests were undertaken at AFDW, 6 RAR and SOI. Because of the small number of females recruited, and the smaller number who qualified to undertake the high intensity tasks, it was not possible to complete a full program of testing in the time available. As planned, most tests were administered twice; however, for reasons discussed in Section 3.2, in a few cases the results of one trial were discarded and a third trial was undertaken. Subject to time constraints, a third trial was also undertaken whenever there was evidence of learning effects.

Table 3. Reliability Testing Program

	Cohort				
Test Name	AFDW	6 RAR	2 RAR	Female	SOI
Simulation tests					
Tests common to Infantry and ADG					
Pack Lift and Place	✓	✓			
Box Lift and Place					✓[1]
Jerry Can Carry	✓	✓		✓	✓
1.82m Wall Climb (running start)	✓	✓		✓	✓
Leopard Crawl	✓	✓		✓	✓
Urban Rushing	✓	✓		✓	✓
Section Attack	✓	✓	✓	✓[2]	✓
Tasks/Tests for ADG only					
Sustained Patrol (5km)	✓				
Pursuit (2.4km)	✓				
Predictive tests					
Tests common to Infantry and ADG					
Loaded Incremental Velocity Run	✓	✓			✓
1.82m Wall Climb (standing start)	✓	✓		✓	✓
Test for Infantry only					
Forced March (10km)		✓		✓[2]	✓[3]

[1] At SOI, the Pack Lift and Place test was replaced by the Box Lift and Place test.
[2] The female soldiers undertook a 10 km Forced March and Section Attack, but only on one occasion. Hence no reliability calculations were possible.
[3] At SOI, the Forced March was 7 km long.

2.5 Measurement of Test Reliability

Statistical Measures of Reliability

2.5.1 Reliability of a test is the extent to which it is consistent and free from error (Portney & Watkins, 1993). Assessing the reliability of a test of physical performance involves testing the same people on two or more occasions over a period of days, to examine the level of agreement (or reproducibility) of test results from different occasions. The spacing between test sessions should be sufficient to allow full recovery from any effects of testing, but not so

long that the effects of maturation, seasonal effects, or the effects of different work or physical training regimens come into play.

2.5.2 When measuring precisely defined and stable quantities such as stature or other anthropometric characteristics, changes in the measured quantity are primarily due to matters of test administration and measurement per se. However, with tests of physical performance, measurement effects are likely to be less important than a range of confounding factors – biological, behavioural and environmental – which lead to variation over time in the characteristic being measured – the performance itself.

2.5.3 Discrepancies between two sets of test scores can be characterised in terms of two components: a shared systematic component which is common to all participants, and an individual component which is specific to each participant.

2.5.4 Three standard measures of reliability for quantitative measurements are intraclass correlation (ICC) (McGraw & Wong, 1996), bias (Bland & Altman, 1986) and limits of agreement (LOA) (Bland & Altman, 1986). The ICC usually takes values between 0 and 1, with 0 indicating no agreement between two (or more) sets of measurements and 1 indicating perfect agreement. The bias (or offset) is the difference between the mean scores of the same group of subjects on two occasions. The LOA is a measure of the expected discrepancy between measurements or scores on the same individual on two occasions (±2 standard deviations[1,2]). Both bias and LOA can be expressed as absolute quantities, or in relative terms as percentages of the mean score on the test.

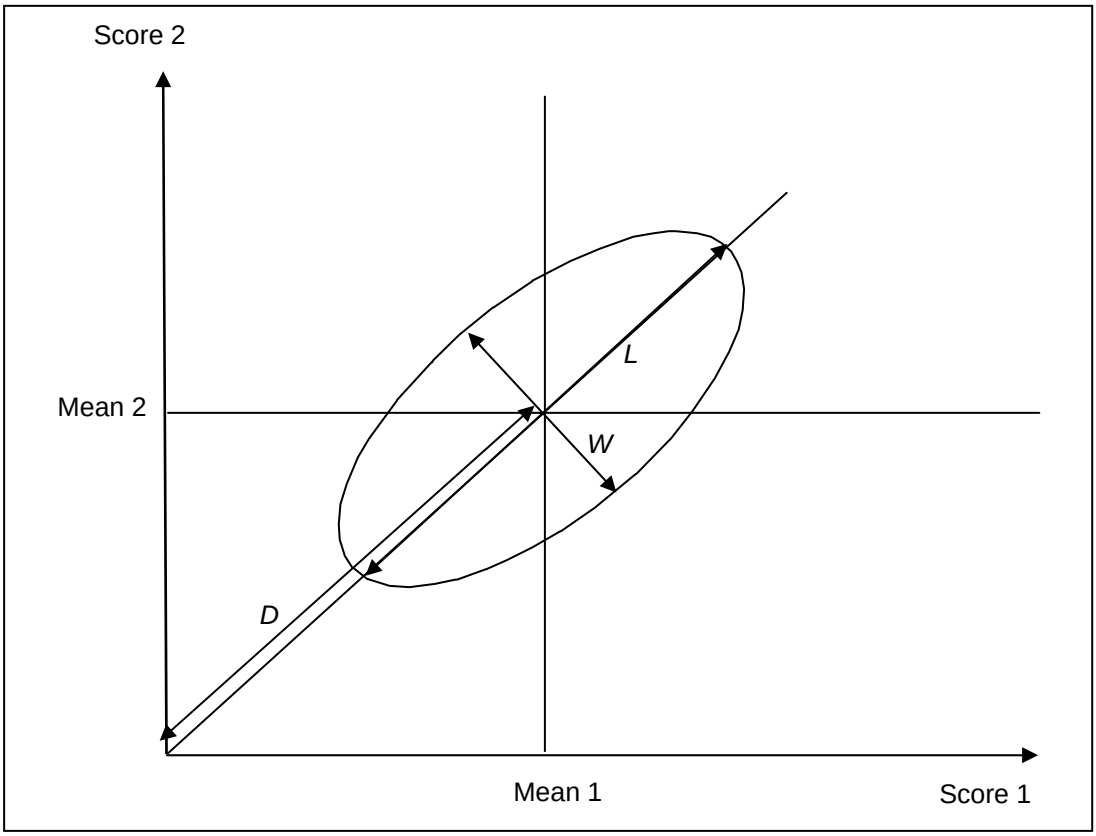


Figure 1. Concepts Underpinning Statistical Measures of Reliability

[1] Some authors use ±1.96 standard deviations

[2] Strictly, Bland & Altman (1986) defined LOA as *bias* ±2 *standard deviations*, which incorporates both common group effects and the overlaid individual variability (as is the case for ICC). Other authors such as Rayson (2004) use LOA to refer to the individual variation only, so that bias and LOA represent two separate aspects of reliability. The latter approach has been adopted in this report.

2.5.5 The ICC is sensitive to both systematic and individual components of difference. Bias is indicative of systematic differences, and LOA is sensitive to changes in the scores of each individual. Two other measures of individual variability (Portney & Watkins (1993) use the term "response stability") are the standard error of measurement (SEM) and the coefficient of variation (CV), which are directly related to the absolute and relative LOA respectively1. These have not been used in this report.

Intraclass Correlation (ICC)

2.5.6 As to what is 'good' reliability, Innes & Straker (1999) point out that since correlation-type measures are very sensitive to sample size, with a large enough sample, almost any correlation coefficient can be found to be statistically significant (i.e. significantly different from zero). But from the perspective of reliability, what is required is not "non-zero reliability" but "high reliability", and hence the magnitude of the coefficient is crucial. Portney & Watkins (1993) suggested that, as a general guideline, an ICC above 0.75 is indicative of good reliability, and that for clinical measurements ICCs should exceed 0.90. Boldovici et al. (2001), writing with reference to military training evaluations, recommended that with "moderate or small samples", test reliability should be 0.70 or greater, and that for decisions about particular individuals or units, reliability should be 0.90 or greater. In an educational context, an acceptable level of ICC > 0.85 has been cited for tests used for critical decisions about individuals (Virginia Department of Education, 1999).

2.5.7 In the context of physical testing of military personnel, Rayson et al. (2004) have pointed out some limitations of the ICC and has de-emphasised its role, preferring instead to use the criterion of relative LOA < 20%, proposed by Prof Alan Nevill (cited in Rayson et al., 2004). However, both ICC and LOA must be interpreted carefully with regard to the context in which they are measured, since both are dependent on the properties of the data. To understand the properties of ICC and LOA, we consider Figure 1, in which two sets of scores are plotted on *X* and *Y* axes. The ellipse schematically represents the "cloud" of data pairs. Assuming for simplicity that there is no substantial systematic overall change (offset or bias) in the scores, then the two means will be similar, and the axis of the ellipse will lie close to a line inclined at 45º to the axes.

2.5.8 The length of the ellipse *L* indicates the range of differences between individuals. It is an indication of how well the test discriminates between individuals. The width of the ellipse *W* relates to differences between the two scores of each individual. The dimension *D* is determined by the mean level of the scores across the group[2]. These three dimensions can each vary independently of the other two.

2.5.9 The ICC indicates the proportion of the variability in the data which is attributable to differences between individuals. Conceptually, it is related to the ratio

$$\frac{L}{L+W}$$

The smaller the discrepancies in the two scores of each individual, the smaller is *W*, and hence the closer the value of the ICC is to 1.

Limits of Agreement (LOA)

2.5.10 The LOA is an absolute measure of consistency, and is conceptually related to the dimension *W*. It is usually expressed in relative terms as a proportion or percentage of the mean, or equivalently

[1] The standard error of measurement (SEM), also referred to by some as Typical Error of Measurement (TEM), is an estimate of the within-subject variability associated with a single measurement. In the case of two sets of measurements, the SEM is equal to the standard deviation of the difference sores divided by √2. Hence LOA = ±2√2(SEM) ≅ 3(SEM). The coefficient of variation (CV) is the SEM expressed as a proportion of the mean. Hence LOA% = ±2√2(CV%) ≅ ±3(CV%).

[2] $D = \sqrt{M_1^2 + M_2^2} \cong \sqrt{2}(M)$ if $M \cong M_1 \cong M_2$

$$\frac{W}{D}$$

The smaller the discrepancies in the two scores of each individual, the smaller is *W*, and hence the closer the value of the both absolute and relative LOA are to 0.

2.5.11 For a given level of test-retest discrepancy *W*, the ICC is sensitive to the value of *L*. If the group tested is a relatively homogeneous subset of the whole population (with respect to the particular aspect of performance being tested), then *L* is reduced (i.e. the test does not discriminate well between the individuals in the group) and hence ICC is reduced; but LOA is unaffected. On the other hand, relative LOA is sensitive to the value of *D*. If the general level of scores is low relative to *W*, then this will tend to increase relative LOA, whilst ICC is unaffected.

2.5.12 It should also be noted that the values of both the ICC and LOA calculated from data are subject to sampling variation. Confidence intervals are routinely calculated for ICCs as for any correlation coefficient. Bland & Altman (1986) demonstrated the calculation of approximate confidence limits for LOA, although these are not generally reported in the reliability testing literature.

Bias

2.5.13 The difference between the means of the scores on two occasions, often referred to as the offset or bias (Bland & Altman, 1986), indicates the "group change" - the extent to which the scores of the group as a whole have changed from one occasion to the next. Like LOA, bias may be expressed in absolute or relative terms.

2.5.14 Bias may indicate changes in the condition of the subjects as a group, brought about by other group activities, or changes in the conditions under which the test is administered. If all such conditions are well controlled, bias may relate to learning effects, individual desire to improve, or the effects of peer pressure and intra-group and inter-group competitiveness. If such group change effects are evident, adequate preparatory familiarisation should be undertaken to reduce bias to a low level before actual (scored) testing takes place. Boldovici et al. (2001) discussed these issues in general terms. Pandorf et al. (2003) provided a specific example of the application of this concept when they demonstrated that two familiarisation trials were necessary to stabilise performance on a 6-station obstacle course, and that one familiarisation trial was sufficient to stabilise performance on a repetitive Box Lift task.

2.5.15 Statistical tests of significance can be performed on the differences between means on successive trials: t-tests for two trials (see for example Pandorf et al., 2003), or analysis of variance for more than two trials. However, as is the case with ICC or LOA, the question as to how small the bias should be cannot be answered by statistical significance alone. Statistical significance is dependent upon sample size, and does not directly indicate the magnitude of the bias, whereas the reporting of estimated bias and associated confidence intervals conveys information about both aspects. When the sample size is small, the tests are not sensitive (i.e. have low statistical power), so that even quite substantial changes in the sample mean may not be statistically significant. Furthermore, many trials may be needed in order to establish a clear pattern and make recommendations regarding the required number of familiarisation trials. For the reasons already outlined (paragraph 2.4.2) it was not always possible in the present study to carry out more than two trials, even if a third trial was indicated. However, in counterpoint to these inherent statistical limitations is an important practical consideration. It is reasonable to assume that when a test developed in the DPESP is adopted by the ADF, candidates will prepare for it and will be supported in their preparation by PTIs. For this reason, it is considered that an analytically-based recommendation regarding the required number of familiarisation trials is not crucial for operational implementation. Adequate familiarisation does however remain an issue for the normative testing data collection phase of the DPESP, when the tests will still be unfamiliar to participants and PTIs.

2.5.16 For categorical data (such as pass/fail data), reliability is assessed by Cohen's kappa statistic, which like ICC is sensitive to both systematic and individual differences. Kappa is calculated from a cross-tabulation of the two sets of categorical scores, and is based on the number of concordant scores (agreements) on two occasions. It is the difference between the observed number of concordances and the number expected under chance allocation, expressed as a proportion of the improvement over chance that is theoretically possible i.e. to perfect agreement. Kappa ranges from 0 (representing the level of agreement expected by chance) to 1 (perfect agreement).

2.5.17 Cohen's kappa has been interpreted as follows: <0.40 poor to fair agreement; 0.41 – 0.60 moderate agreement; 0.61 – 0.80 substantial agreement; >0.80 excellent to perfect agreement (Innes & Straker, 1999).

2.5.18 Inference about kappa is not straightforward. Because kappa is essentially a proportion, its sampling distribution becomes increasingly negatively skewed as kappa approaches 1. The sampling behaviour of kappa is also dependent on the prevalence of the condition (in this context, the proportion of candidates who pass the test). For large samples (>30), a broad indication of sampling variability can be given by an approximate lower confidence limit based on an approximate (asymptotic) standard error.

Inter-tester Reliability and Intra-tester or Test-retest Reliability

2.5.19 In physical tests which are administered and scored by an individual tester or a team of testers it is important to establish both intra-tester reliability (the same tester on two or more occasions)[1] and also inter-tester reliability (different testers on each occasion). The latter is particularly important for establishing generalisability (Portney & Watkins, 1993; Innes & Straker, 1999), which is crucial for operational field-based testing. It can be argued that intra-tester reliability is a necessary pre-condition for inter-tester reliability, and hence that establishing the latter implicitly also establishes the former. In the DPES project, there were insufficient PTIs available at each location for inter-tester reliability to be assessed in a pure and unconfounded manner. In most cases, the functions of individual members of the testing team were rotated between trials to the degree that this was feasible. Hence the reported reliabilities represent an amalgam of inter-tester and intra-tester reliability.

Standards of Reliability Adopted in this Study

2.5.20 In this report, reliability of quantitative measures is assessed on the basis of ICCs calculated using two-way random effects models to account for differences between subjects and between administrations of the test, and testing for absolute agreement between the test scores on each administration (McGraw & Wong, 1996). Bias and LOA are also considered. For categorical data, Cohen's kappa is used.

2.5.21 The results for each cohort are indicative; the sample sizes at each location were so small that statistical confidence bounds calculated for ICCs and changes in the mean (biases) were generally very wide, indicating a wide range of uncertainty in the sample estimates. On the one hand, this meant that definitive conclusions about reliability in particular cohorts were not possible; on the other hand, it meant that even the largest biases were not statistically significant i.e. not conclusively demonstrated to exist. In some cases, particular circumstances prevailing at different locations limited the validity of pooling results across cohorts/locations in order to achieve the sample sizes specified in the study design. However for each test, results which are considered sufficiently valid and comparable have been pooled. Confidence intervals have been reported for values of ICC, bias, and LOA calculated from pooled data.

2.5.22 In the light of the foregoing discussion, target levels adopted in this report for three of the four reliability indicators are: ICC>0.85; LOA< 20%; and kappa>0.60. Having failed to find any recommendation in the literature with regard to an acceptable magnitude (as opposed to statistical significance) of bias, a target of bias<5% was adopted. These are not regarded as

[1] Intra-tester reliability is a special case of test-retest reliability.

inflexible and prescriptive requirements, but rather as indicative guides to an acceptable level of reliability.

2.6 Designing for Reliability

2.6.1 Boldovici et al. (2001) listed a set of 18 rules for eliciting reliable ratings for the purpose of individual and collective performance appraisal. Whilst not all of the rules are relevant to the context of physical testing, this list provided a useful benchmarking tool when planning and implementing the reliability testing program. The full set of rules, together with comments regarding applicability and adoption, are listed in Table 4.

Table 4. Rules for Eliciting Reliable Ratings

	Phase/Rule	**Comment**
	Scope of reliability assessments	
1	Estimate and report inter-rater reliability and its implications for validity.	See paragraph 2.5.19
2	Estimate and report generalisability.	See paragraph 2.5.19
	Preparation of Raters	
3	Be specific in instructions to raters.	Adopted (See protocols Annex 9)
4	Provide instructions early enough to allow practice, feedback, and learning.	Adopted – WO PTIs engaged early – PTIs recruited early
5	Provide practice in observing and rating.	Only possible “on the day”
6	Test raters.	Not feasible
	Observation	
7	Deconstruct multi-dimensional criteria	Adopted (See protocols Annex 9)
8	Deconstruct multi-dimensional events.	Adopted (See protocols Annex 9)
9	Make transient events stable (e.g. by filming then rating later)	All testing was videotaped; in one case the tapes were used in the review of a scoring protocol.
10	Avoid noise in rated events.	Adopted e.g. CATT simulation tests on oval rather than in bush
11	Strive for observability in rated events.	Not relevant
12	Require comparative rather than absolute judgments	Not relevant
13	Alert raters to likely errors.	Adopted
14	Allow raters to observe and rate more than once.	Not feasible
15	Provide scoring aids or templates.	Not relevant
16	Do not require raters to process results.	Adopted
	Recording	
17	Keep the time short between observing and recording.	Adopted
18	Keep rating forms simple.	Adopted (see sample recording form Annex 10)

3 RESULTS

3.1 Outcome Measures

3.1.1 The outcome measures for each test are listed in Table 5. Most tests were too physically demanding and fatiguing to perform more than once in a session. However in three cases - the Leopard Crawl, the Wall Climb and the Section Attack[1] - multiple performances of the basic test procedure in a single session were feasible, which provided some flexibility with regard to the test protocols and the outcome measures for these three tests. Final decisions on these tests were made with regard to the results of the reliability analysis. The basis of these decisions is discussed in detail in Section 4.

3.1.2 *Leopard Crawl*. Two attempts were undertaken, separated by 10 min. Four duration measures were investigated: first attempt, second attempt, mean of two attempts, best of two attempts. The best of two attempts exhibited the best reliability, and was adopted as the measure for this test.

3.1.3 *Wall Climb*. Each attempt consisted of a climb with a running start followed by reverse climb from a standing start. Three attempts were undertaken, separated by 10 min. Hence a success/failure score, and in the case of success, a duration, were available for each of three attempts with running starts and three attempts with standing starts. The reliability of many different composite measures were tested. The measures adopted were as follows:

Categorical. Three measures were adopted: pass/fail on first attempt; pass on all three attempts v. fail on some or all attempts; fail on all three attempts v. pass on some or all attempts.

Quantitative. Fastest time from three attempts.

3.1.4 *Section attack*. Three duration measures were investigated: first (outward) leg, second (return) leg, best of two legs, and total for both legs. The total time for both legs was adopted as the preferred measure for this test.

Table 5. Test Measures

Test Name	Measure	Unit
Simulation tests		
Tests common to Infantry and ADG		
Pack Lift and Place	Number of packs loaded	Number
Box Lift and Place	Weight of heaviest box lifted	15 kg – 45 kg in 5 kg increments
Jerry Can Carry	Distance carried	m
1.82m Wall Climb (running start)	Completion in set time	Pass/fail
	Time taken (best of three)	sec
Leopard Crawl	Time taken (best of two)	sec
Urban Rushing	Stages completed	Number
Section Attack	Total time taken (out & back)	sec
Tasks/Tests for ADG only		
Sustained Patrol (5km)	Completion in set time	Pass/fail
Pursuit (2.4km)	Completion in set time	Pass/fail
Predictive tests		
Tests common to Infantry and ADG		
1.82m Wall Climb (standing start)	Completion in set time	Pass/fail
	Time taken (best of three)	sec
Loaded Incremental Velocity Run	Level & stage reached	Decimalised level/stage
Test for Infantry only		
Forced March (10km)	Completion in set time	Pass/fail

[1] The section attack was only conducted once per trial but times were recoded for “out” and “return” legs of the course, which meant that there three options could be considered for a test score: the two separate times and the combined time.

3.2 Results of Reliability Analyses

3.2.1 The results of the reliability analyses on categorical data are presented in Table 6.

Table 6. Results of Reliability Testing: Categorical Data

Test	Unit	Trials	Initial sample size	Final sample size *n*	Success/ failure on first attempt kappa[1 2 3 4]	Succeeded on all attempts v. not all kappa[1 2 3 4]	Succeeded on no attempts v. some kappa[1 2 3 4]
Wall Climb (Running)	AFDW	1 & 2	7	7	NC+	NC+	NC+
	6 RAR	1 & 2	9	7	NC+	NC+	NC+
		2 & 3	7	3	NC+	NC+	NC+
		1 & 3	9	3	NC+	NC+	NC+
	Female	1 & 2	10	8	0.60	0.60	NC+
	SOI	1 & 2	8	8	0.71	0.33	0.53
		2 & 3	8	8	0.60	0.60	0.71
		1 & 3	8	8	0.39	0.60	0.33
	Pooled	1 & 2	34	30	0.71 (LCL≅0.33)	0.67 (LCL≅0.33)	0.46 (LCL≅0)
		2 & 3	15	11	0.62	0.74	0.62
		1 & 3	17	11	0.42	0.42	0.62
Wall Climb (Standing)	AFDW	1 & 2	7	7	NC+	NC+	NC+
	6 RAR	1 & 2	9	7	1.00	NC+	1.00
		2 & 3	7	3	NC+	NC+	NC+
		1 & 3	9	3	NC+	NC+	NC+
	Female	1 & 2	10	8	NC-	NC-	NC-
	SOI	1 & 2	8	8	1.00	0.71	1.00
		2 & 3	8	8	0.25	0.60	0.25
		1 & 3	8	8	0.25	0.39	0.25
	Pooled	1 & 2	34	30	0.63 (LCL≅0.26)	0.93 (LCL≅0.79)	0.79 (LCL≅0.55)
		2 & 3	15	11	0.21	0.27	0.42
		1 & 3	17	11	0.21	0.27	0.42
Sustained Patrol	AFDW	1 & 2	8	7	NC+	NA	NA
Pursuit	AFDW	1 & 2	10	6	NC+	NA	NA
Forced March (10 km)	6 RAR	1 & 2	8	4	NC+	NA	NA
	SOI	1 & 2	8	8	NC+	NA	NA

[1] NC+ = not calculated (consistent positive outcomes). In these cases, on at least one occasion all participants passed the test (i.e. completed the activity within the required time range). Kappa cannot be calculated unless there are passes and fails on both occasions.

[2] NC- = not calculated (consistent negative outcomes). In these cases, on at least one occasion all participants failed the test (i.e. failed to complete the activity within the required time range). Kappa cannot be calculated unless there are passes and fails on both occasions.

[3] NA = not applicable.

[4] LCL = approximate 95% lower confidence limit.

3.2.2 For a test with a dichotomous outcome (i.e. two categories – pass and fail), the reliability of a test indicates the capacity of the test to consistently discriminate between a pass and fail standard. If all participants who undertake the test pass it (i.e. satisfactorily complete the activity within the required time frame) either in one trial or in both trials, it is impossible to calculate kappa or to assess the reliability of the test, because there is insufficient evidence available with regard to discrimination. Table 6 shows that this occurred for all three extended endurance tasks - the Sustained Patrol and Pursuit (ADG) and the Forced March (Infantry) – in which all participants who took part in both trials passed on both occasions.

3.2.3 For the two Wall Climb tests (with running start and standing start), three dichotomous measures were considered in each case:

- Pass/fail on the first attempt
- Pass on all three attempts v. fail on some or all attempts
- Fail on all three attempts v. pass on some or all attempts

The results are shown in Table 6. In a number of the analyses, all participants passed in either one trial or both trials. In the case of females with standing starts, all participants failed in either one trial or both trials.

3.2.4 The results of reliability analyses on quantitative data are presented in Table 7.

3.2.5 The Wall Climb test can be of particularly short duration, and so in this test accurate timekeeping is particularly crucial. Two timekeepers were used in each trial and the mean of their times was used. Using pooled data from all attempts and all trials, for each pair of timers all measures indicated a high level of inter-rater reliability (ICC>0.99, Bias<2% and LOA<8%).

Table 7. Results of Reliability Testing: Quantitative Data

Test	Unit	Trials	Initial sample size	Final sample size *n*	Intraclass correlation[1] ICC	Unit of meas.	Mean	Change in mean[1] (Bias)	Relative Bias %	Limits of agreement[1] (LOA)	Relative LOA %	Notes
Pack Lift & Place	AFDW	1 & 2	9	9	0.15	kg	15.16	+0.11	+1%	±20.22	133%	
	6 RAR	1 & 2	6	4	0.61	kg	12.75	+3.50	+27%	±16.45	129%	
Box Lift & Place	SOI	1 & 2	8	8	0.74	kg	37.82	+1.88	+5%	±5.18	14%	
		2 & 3	8	8	0.58	kg	37.19	-3.13	-8%	±7.44	20%	
		1 & 3	8	8	0.84	kg	36.26	-1.25	-3%	±4.63	13%	
					(0.43,0.97)			(-3.19,+0.69)		±(1.36,7.90)		
Jerry Can Carry	AFDW	1 & 2	9	9	0.88	m	258.8	+20.5	+4%	±62.0	24%	
	6 RAR	1 & 2	7	7	0.25	m	302.9	+52.9	+17%	±265.8	103%	
		2 & 3	8	8	0.97	m	334.0	+6.9	+2%	±76.3	23%	
	Female	1 & 2	6	6	0.69	m	78.5	-7.1	-9%	±58.7	75%	
		2 & 3	6	5	0.32	m	95.3	+26.6	+28%	±86.8	91%	
		1 & 3	6	5	0.81	m	95.3	+18.6	+20%	±31.8	33%	
	SOI	1 & 2	8	7	0.97	m	451.1	+10.7	+2%	±71.3	16%	
	Pooled			24	0.96	m	336.0	+5.5	+2%	±77.6	23%	AFDW 1&2; 6 RAR 2&3; SOI 1&2; Female omitted
					(0.92,0.98)			(-10.9,+21.9)		±(49.2,106.0)		
Wall Climb (Running)	AFDW	1 & 2	7	7	0.79	s	4.12	+0.25	+6%	±0.50	12%	
	6 RAR	1 & 2	8	7	0.49	s	4.67	-0.07	-1%	±0.92	20%	
		2 & 3	7	3	0.32	s	4.65	+0.24	+5%	±1.50	32%	
		1 & 3	8	3	0.66	s	4.77	-0.02	-0.4%	±1.28	27%	
	Female	1 & 2	10	8	0.88	s	9.34	-0.19	-2%	±2.72	29%	
	SOI	1 & 2	8	5	0.02	s	6.98	-0.40	-6%	±2.42	35%	
		2 & 3	6	6	0.86	s	7.14	+0.06	+1%	±1.14	16%	
	Pooled			28	0.96	s	6.39	<0.01	<1%	±1.59	25%	AFDW 1&2; 6 RAR 1&2; Female 1&2; SOI 2&3
					(0.91,0.98)			(-0.31,+0.30)		±(1.06,2.12)		
Leopard Crawl	AFDW	1 & 2	10	8	0.91	s	25.55	-0.57	-2%	±3.27	13%	
	6 RAR	1 & 2	9	8	0.82	s	24.82	-0.78	-3%	±5.34	22%	
		2 & 3	9	7	0.82	s	25.45	+1.38	+5%	±5.18	20%	
		1 & 3	9	7	0.77	s	26.07	+0.13	+1%	±6.54	25%	
	Female	1 & 2	6	5	0.81	s	52.23	-6.17	-12%	±12.38	24%	
		2 & 3	5	4	0.93	s	45.53	+1.26	+3%	±5.97	13%	
		1 & 3	6	5	0.71	s	50.83	-6.35	-12%	±12.99	26%	
	SOI	1 & 2	8	8	0.27	s	27.54	-3.55	-13%	±6.64	24%	
		2 & 3	8	6	0.32	s	26.20	-0.27	-1%	±6.68	25%	
		1 & 3	8	6	0.60	s	28.00	-3.79	-14%	±1.98	7%	
	Pooled			26	0.95	s	28.54	-0.28	-1%	±5.11	18%	AFDW 1&2; 6 RAR 1&2; Female 2&3; SOI 2&3
					(0.91,0.98)			(-1.32,+0.75)		±(3.33,6.89)		

[1] 95% confidence limits in parentheses

Table 7. Results of Reliability Testing: Quantitative Data (Continued)

Test	Unit	Trials	Initial sample size	Final sample size n	Intraclass correlation[1] ICC	Unit of meas.	Mean	Change in mean[1] (Bias)	Relative Bias %	Limits of agreement[1] (LOA)	Relative LOA %	Notes
Urban Rushing	AFDW	1 & 2	9	9	0.78	lap	28.28	+1.89	+7%	±12.70	45%	
	6 RAR	1 & 2	9	6	0.65	lap	27.17	-0.78	-4%	±5.34	22%	
	Female	1 & 2	4	3	0.61	lap	12.83	+1.00	+8%	±4.48	35%	
		2 & 3	3	3	0.61	lap	14.67	+3.00	+20%	±2.00	14%	
		1 & 3	4	3	0.44	lap	14.33	+4.00	+28%	±4.00	28%	
	SOI	1 & 2	8	5	0.83	lap	39.30	+6.60	+17%	±8.08	21%	
	Pooled			23	0.84 (0.64,0.93)	lap	28.63	+2.74 (-0.08,+5.40)	+10%	±12.28 ±(7.69,16.87)	43%	AFDW 1&2; 6 RAR 1&2; Female 2&3; SOI 1&2
Section Attack (out and back)	2 RAR	1 & 2	9	8	0.85	s	84.19	+0.63	+1%	±9.80	12%	
	AFDW	1 & 2	9	5	0.80	s	117.97	+0.52	+1%	±25.52	22%	
	6 RAR	1 & 2	8	4	-0.43	s	106.24	+5.64	+5%	±45.06	42%	
	SOI	1 & 2	8	8	0.71	s	104.80	-14.68	-14%	±15.30	15%	
	Pooled			21	0.84 (0.62,0.93)	s	100.08	-5.23 (-10.20,-0.26)	-5%	±21.84 ±(13.24,29.44)	22%	6 RAR omitted
Section Attack (out only)[3]	AFDW	1 & 2	9	5	0.68	s	51.43	-1.82	-4%	±12.06	23%	
	6 RAR	1 & 2	8	4	-0.59	s	44.48	+5.33	+12%	±19.22	43%	
	SOI	1 & 2	8	8	0.66	s	43.50	-5.62	-13%	±7.28	17%	
	Pooled			13	0.72	s	46.55	-4.16 (-7.09,-1.22)	-11%	±9.72 ±(5.05,14.38)	21%	6 RAR omitted
Wall Climb (Standing)	AFDW	1 & 2	7	7	0.88	s	3.80	+0.13	+3%	±1.20	32%	
	6 RAR	1 & 2	8	5	0.67	s	4.96	-0.53	-11%	±2.68	54%	
		2 & 3	5	2	-8.78	s	4.22	+0.06	+1%	±1.36	32%	
		1 & 3	8	3	0.83	s	5.45	+1.16	+21%	±2.54	47%	
	Female	1 & 2	10	8	NC[2]	s	NC[2]	NC[2]	NC[2]	NC[2]	NC[2]	
	SOI	1 & 2	8	5	-0.10	s	6.58	-0.85	-13%	±1.72	26%	
		2 & 3	6	6	0.79	s	6.97	+.70	+10%	±2.11	30%	
Loaded Incr. Velocity Run	AFDW	1 & 2	9	9	0.62	level	7.70	+0.66	+8%	±1.44	19%	
	6 RAR	1 & 2	8	7	0.66	level	7.15	-0.38	-5%	±1.42	20%	
	SOI	1 & 2	9	6	0.89	level	8.85	-0.08	-1%	±1.67	19%	
	Pooled			22	0.80 (0.57,0.91)	level	7.84	+0.13 (-0.24,+0.51)	+2%	±1.73 ±(1.06, 2.40)	22%	

[1] 95% confidence limits in parentheses
[2] NC = not calculated. In these cases, on at least one occasion no participants passed the test (i.e. completed the activity within the required time range).
[3] Times for the outward leg of the Section Attack were not recorded in the preliminary testing at 2 RAR.

4 DISCUSSION

4.1 Limitations

4.1.1 In tests on which performance includes a skill or confidence component, reliability is improved by making adequate provision for familiarisation. On the advice of Defence informants, it was assumed that all participants would be familiar with the commonly performed CATTs on which the simulation tests were based. Whilst this was the case with the participants in the pilot testing, it became apparent during reliability testing that in most cases participants would have benefited from more extensive opportunity for familiarisation than was possible within the time constraints of the DPESP reliability field trials. Tests for which some cohorts particularly require familiarisation are indicated.

4.1.2 It is a commonly held view that reliability is a necessary precursor to validity – that a test which is not reliable cannot be said to be valid, and hence should not be used for making clinical or other crucial decisions. It can certainly be shown (for example Boldovici et al., 2001) that at a technical level, there is a direct relationship between quantitative measures of reliability and predictive validity, and that the latter cannot be greater than the former. However, this is not the case with respect to criterion validity. On the contrary, it has been pointed out that there is a trade-off between criterion validity and reliability; "as the test situation simulates reality more closely, control becomes more difficult…the more closely one tries to simulate the criterion situation, the less reliable will be one's measurement of performance" (Fitzpatrick & Morrison, 1971, cited in Rosen, 1978). Since the rationale of the DPESP project is very much focused on criterion validity and task-related simulation tests, this limits the levels of reliability that could reasonably be expected.

4.1.3 The results are indicative; the sizes of the samples made available by Defence at each location were so small that statistical confidence bounds calculated for ICCs and LOAs were generally too wide (and hence uninformative) to be worth reporting, and even very large biases (changes in the mean) were not statistically significant. In some cases, particular circumstances prevailing at different locations limited the validity of pooling results across cohorts/locations in order to achieve the sample sizes specified in the study design. However for each test, results which are considered sufficiently valid and comparable have been pooled. Confidence intervals have been reported for values of ICC, bias, and LOA calculated from pooled data.

4.1.4 According to the standards discussed in Section 2.5, the levels of reliability obtained are generally only marginally acceptable. This is considered to be, at least in part, a consequence of the conditions under which testing took place. The reliability of a test is maximised when the test sample is drawn from a relatively homogeneous population, and the condition of the subjects and the conditions under which the test is administered are tightly controlled, as in a laboratory testing environment, so that the effects of extraneous sources of variation are minimised. Whether the levels of reliability measured under such pristine experimental conditions are ever attained in operational conditions is a moot point. Be that as it may, the conditions under which reliability was assessed in the DPES project were challenging. The scope of the reliability testing program was ambitious, considering the number of tests to be evaluated, the range of personnel to be tested, and the limited time and numbers of personnel available at each location. The test sample was drawn from four populations which differed with regard to skill and experience, general physical capability, level of fitness, work cycles, prior injury status and motivation. The attitude and level of commitment of PTIs, and the thoroughness of their compliance with test protocols, differed at different locations, as did the level of support, encouragement and commitment of officers.

4.1.5 On a more positive note, from a practical perspective the results of this investigation arguably represent a robust and realistic assessment of reliability under "worst case" in-field operational conditions. Furthermore, the tight DPESP schedule meant that the spacing of trials of different tests was close enough that residual effects of prior tests could not be ruled out, and there was limited opportunity for familiarisation and preparation for both participants and PTIs. It is considered that if and when these tests are adopted and become well-established and familiar, and when candidates prepare for the tests over a longer period, the reliability will increase. Thus the reliabilities reported represent a "floor" or worst case level.

4.2 Discussion of Categorical Results

Tests Common to Infantry and ADG

4.2.1 *1.82m Wall Climb (running and standing starts).* As discussed previously (paragraph 3.1.3), these tests were assessed using both categorical and quantitative measures. Three dichotomous measures were considered:

- Pass/fail on the first attempt
- Pass on all three attempts v. fail on some or all attempts
- Fail on all three attempts v. pass on some or all attempts

4.2.2 As discussed previously (paragraph 3.2.3), it was generally not possible to calculate the standard measure of reliability (kappa) for these tests as applied to experienced soldiers and airmen, since in most trials at AFDW and 6 RAR all participants passed on both occasions. Conversely, all females tested failed the wall climb for a standing start. However, results were obtained for IETs at SOI and for pooled data from all four cohorts.

4.2.3 Considering the first two trials in each case, for both running start and standing start the agreement of success/failure at the first attempt in each trial was, in the terminology of Innes & Straker (1999), substantial (kappa for pooled data = 0.71 and 0.63 respectively). In the case of the running start, the agreements for the two measures based on all three attempts were both lower, being substantial (kappa = 0.67) and moderate (kappa = 0.46) respectively. In the case of the standing start, the reliabilities for the two measures based on all three attempts were both higher (kappa for pooled data = 0.93 and 0.79 respectively). However, the differences were not statistically significant at the 5% level[1]. No improvement in reliability was found when the results of a third trial in two of the cohorts were considered[2].

4.2.4 It is concluded that success/failure on the first attempt is a reliable measure for Wall Climb tests from both running and standing starts. In the case of the standing start, there is some indication that the more complex dichotomies "pass on all three attempts v. fail on some or all attempts" and "fail on all three attempts v. pass on some or all attempts" might be more reliable measures of success/failure.

Tests Specific to Infantry or ADG

4.2.5 *Forced March (10km) (Infantry).* This test is administered as a pre-fatiguing activity prior to the Section Assault, and is scored as a categorical pass/fail. To pass the test, the participant must complete the course at the prescribed pace and in the prescribed time.

4.2.6 *Sustained Patrol (5km) (ADG)*. This test is scored as a categorical pass/fail. To pass the test, the participant must complete the course at the prescribed pace and within the prescribed time.

4.2.7 *Pursuit (2.4km) (ADG)*. This test is administered as a pre-fatiguing activity prior to the Section Assault, and is scored as a categorical pass/fail. To pass the test, the participant must complete the course at the prescribed pace and in the prescribed time.

4.2.8 As discussed previously (paragraph 3.2.2), it was not possible to calculate the standard measure of reliability (kappa), for any of these tests, because all participants passed the tests on both occasions. However, these tests are categorical "hurdles" which it is expected that all IE trainees and trained soldiers/ADGs should be able to pass. Thus, in practice, consistent discrimination between positive and negative outcomes - usually a key aspect of test reliability - is not an issue with these tests. The tests were demonstrated to be reliable in

[1] It should be noted that the statistics used for assessing categorical measures are based on proportions, which inherently exhibit large sampling variation. Hence, a large sample size (in the order of hundreds) is necessary to conclusively establish a difference in the order of 0.20 (20 percentage points) in a sample proportion such as a kappa statistic.

[2] In fact in each case the kappa based on trials 2 and 3 was lower. However, these results were based on a much smaller sample and are not regarded as comparable or reliable.

the positive sense that all participants achieved the same positive outcome on both occasions.

4.3 Discussion of Results for Quantitative Measures

4.3.1 All tests involving quantitative measures are common to both Infantry and ADG. These include seven simulation tests and two predictive tests.

Simulation Tests

4.3.2 *Pack Lift and Place*. Issues of safety and reliability emerged in pilot testing (see paragraph 2.3.13) and modifications were made in an attempt to overcome the problems. However, in the reliability testing program it became apparent that problems remained, and ultimately the test was rejected on both reliability and safety grounds. The score on this test is the number of packs lifted and placed whilst maintaining a set pace and correct technique and form. It proved difficult to standardise the pack types, pack adjustments, pack contents, weight distribution, the manner of holding and lifting the packs, and hence it was also difficult to standardise judgments about good form and stopping criteria. All reliability measures were very poor, at both AFDW and 6 RAR. It was decided to omit this test from the reduced female testing program, and subsequently, to discard the test altogether and replace it by the Box Lift and Place test.

4.3.3 *Box Lift and Place*. This test, which is closely modeled on a test used in the UK (Ministry of Defence, undated), replaced the Pack Lift and Place, and was tested only at SOI. Three trials were run. The score on this test is the maximum weight lifted and placed whilst maintaining a set pace and correct technique and form. On the second trial, some participants changed their technique to a more continuous "clean" style of lift, which enabled them to lift heavier weights. The protocol was clarified and more tightly enforced in the third trial. For this reason, data from trial 2 was discarded. The reliability measures obtained from trials 1 and 3 (ICC = 0.84, significant[1]; Bias = -3%, not significant; LOA = 13%) are regarded as acceptable.

4.3.4 *Jerry Can Carry*. The score on this test is the distance carried whilst maintaining a set pace and correct technique and form. ICC and bias measures obtained at AFDW were good and LOA marginally acceptable (ICC = 0.88; Bias = +4%; LOA = 24%). The first trial at 6 RAR was characterised by poor motivation and variable degree of effort by participants. These problems were addressed and remedied in trial 2. As a result, reliability measures based on trials 1 and 2 were very poor ((ICC = 0.25; Bias = +17%; LOA = 103%). Results for trials 2 and 3 were similar to those at AFDW (ICC = 0.97; Bias = +2%; LOA = 23%). All three measures obtained at SOI were good (ICC = 0.97; Bias = +2%; LOA = 16%). An acceptable level of reliability was not established for females. Three trials were run, the first two with six participants and the third with five participants. The three results for one of the five participants were very variable, which impacted strongly on all measures of reliability. The best levels of reliability were achieved between trials 1 and 3, but both bias and LOA remained unacceptably high (ICC = 0.81; Bias = +20%; LOA = 33%). It is considered that an acceptable level of reliability for this test has been established for males but not for females. The pooled results for 24 males (trials 1 & 2 at AFDW and SOI; trials 2 & 3 at 6 RAR) were: ICC = 0.96, significant; Bias = +2%, not significant; LOA = 23%.

4.3.5 *1.82m Wall Climb (running start).* As discussed above (paragraph 3.1.3), this test is assessed using both categorical and quantitative measures. Participants are allowed three attempts to scale a 1.82m wall with a running start. The time limit for each attempt is 30 seconds, but if the participant falls back to the ground the attempt is terminated and scored as a fail.[2] For successful attempts, a number of quantitative measures (all times) were considered, including:

[1] Throughout Section 4.3, "significant" is shorthand for "significantly different from zero at the 5% level". A result is significant when the confidence interval in Table 6 does not include zero. Significance testing with respect to zero is meaningful for ICC (where significance is desired) and bias (where it is not), but not for LOA.

[2] Initially, multiple attempts were permitted within each 30 second period ("attempts within attempts"). However, after some initial analysis, it was decided that more reliable time scores would be obtained if multiple attempts were not permitted. This is

- time taken on the first attempt
- time taken on the second attempt
- time taken on the third attempt
- mean time for all successful attempts
- fastest (minimum) time of three attempts.

4.3.6 The quantitative measure finally selected, on which the results in Table 7 are based, was the fastest time recorded from three attempts. This was consistently much more reliable than any of the individual times and marginally more reliable than the mean of the three times. Therefore it was adopted as the measure for this test.

4.3.7 The LOA result from AFDW was good (12%) and results for ICC and bias were marginally acceptable (ICC = 0.79; Bias = 6%). At 6 RAR, ICC was low but bias and LOA were acceptable (ICC = 0.49; Bias = -1%; LOA = 20%). Females exhibited high ICC and low bias but unacceptably high LOA (ICC = 0.88; Bias = -2%; LOA = 29%). At SOI, reliability results from the first two trials were not good (the very low ICC in particular being exacerbated by the homogeneity of the group – see paragraph 2.5.11), but results from trial 2 and trial 3 were acceptable (ICC = 0.86; Bias = +1%; LOA = 16%). These results are regarded as acceptable for male soldiers and airmen and for well-familiarised IE trainees, and marginally acceptable for female soldiers. The pooled results for 20 males (trials 1 & 2 at AFDW and 6RAR; trials 2 & 3 at SOI) and 8 females (trials 1 & 2) were: ICC = 0.96, significant; Bias <1%, not significant; LOA = 25%.

4.3.8 *Leopard Crawl*. The score on this test is the time to complete a set 25 m course. Because the test is short and not very fatiguing, it was decided to allow two attempts in each trial and investigate four potential measures: time for first attempt; time for second attempt; combined time for both attempts; and time for best attempt. The time for the best attempt proved to be consistently more reliable than any of the other times (though not necessarily so with respect to every reliability measure in every cohort on every occasion), and hence it was adopted as the measure for this test.

4.3.9 Very good reliability measures were obtained at AFDW (ICC = 0.91; Bias = -2%; LOA = 13%). Results from the first two trials at 6 RAR were marginally acceptable (ICC = 0.82; Bias = -3%; LOA = 22%), with no further improvement at trial 3. Female results were similar except for a more substantial bias, perhaps due to learning effects, from trial 1 to 2 (ICC = 0.81; Bias = -12%; LOA = 24%). Results for trials 2 and 3 were much better, with the bias greatly diminished and improvements in ICC and LOA also (ICC = 0.93; Bias = -3%; LOA = 13%). IE trainees at SOI also exhibited a group change effect (Bias = -12%) from trial 1 to trial 2. Whilst the bias was reduced to -1% at trial 3, the LOA remained unacceptably high at 25%. ICC remained very low throughout, reflecting the fact that this group was very homogeneous, with very similar scores on this particular task (see paragraph 2.5.11). In what may be no more than a random occurrence, the best values of ICC and LOA (0.60 and 7%) occurred between trials 1 and 3, while the lowest bias occurred between trials 2 and 3. Overall, these results indicate acceptable reliability of this test for experienced combat arms personnel, with the necessity for some preliminary familiarisation for females and IE trainees. The pooled results for 22 males (trials 1 & 2 at AFDW and 6RAR; trials 2 & 3 at SOI) and 4 females (trials 2 & 3) were: ICC = 0.95, significant; Bias -1%, not significant; LOA = 18%.

4.3.10 *Urban Rushing*. The score on this test is the number of laps of a 22 m course covered whilst maintaining a set pace. Reliability measures obtained at AFDW were unacceptable (ICC = 0.78; Bias = +7%; LOA = 45%); there was no apparent reason for this. Whilst the ICC obtained at 6 RAR was lower (0.65), both bias (-4%) and LOA (22%) were more acceptable (indicating that the low ICC was probably due to homogeneity of performances rather than unreliability per se – see paragraph 2.5.11). A very small number of females undertook three trials of this test, with very poor results. ICCs ranged from 0.44 to 0.60, and there was a steady increase in the mean scores from trial 1 to trial 3, with the bias in each case being beyond acceptable limits. Only in one instance was the LOA within acceptable limits. Two

because someone who succeeds on the first attempt might take say 7 sec. On a subsequent occasion, he/she might miss his/her footing or hit the wall awkwardly, and hence miss at the first attempt, return to the starting point and subsequently succeed on the second jump in say 8 seconds, but with a recorded total elapsed time of 16 seconds – a very different time from the first occasion. All the data were re-analysed using the single attempt protocol.

trials at SOI produced the highest ICC (0.83) and a marginally acceptable LOA (22%), but a substantial bias (+17%), which once again suggested the possibility of a learning effect on this CATT- based test among IE trainees. These results indicate that the reliability of this test is potentially acceptable for well-familiarised IE trainees, and marginally acceptable for trained soldiers. The results for females suggest that an extensive program of familiarisation would be required in order for this test to be reliable for female soldiers. The pooled results for 20 males (trials 1 & 2 at AFDW, 6RAR and SOI) and 3 females (trials 2 & 3) were: ICC = 0.84, significant; Bias +10%, not significant; LOA = 43%. The high LOA in the pooled results is due to a combination of high LOA in the AFDW results and the mixture of positive and negative biases in the individual cohorts.

4.3.11 *Section Attack*. The score on this test was the time to complete a prescribed assault course. Potential measures investigated were: time for 'outward' leg; time for return leg; combined time for both legs; and time for best leg. The total time for both legs (out and back) was more reliable than the other times in most respects and most instances, and hence it was provisionally adopted as the measure for this test. The results discussed below are based on this measure. However, the difference in the reliability of the time for the outward leg and the total time was not great, and it is recognised that a test based on a single leg of the course may be preferable on other grounds (such as duration of the test; face validity and acceptance by soldiers and airmen; logistics of test administration). For this reason, results based on the time for the outward leg only are also included in Table 7 for further consideration if required.

4.3.12 Reliability measures obtained at AFDW were marginal (ICC = 0.80; Bias = +1%; LOA = 22%). The ICC and LOA obtained at 6 RAR were unacceptable (ICC = -0.43; Bias = +5%; LOA = 42%). This was due in large measure to the pre-existing physical condition of the participants, who generally found the pre-fatiguing 10 km forced march very challenging. Four of the eight participants did not take part in the second trial. The female soldiers undertook a 10 km Forced March and Section Attack, but only on one occasion. Hence no reliability calculations were possible. IE trainees at SOI produced relatively homogeneous sets of scores resulting in a low ICC (0.71), but with an acceptable LOA (15%). A substantial bias (-14%), again suggested the possibility of a learning effect on this CATT- based test among IE trainees. These results were supplemented by the results of some preliminary testing in May 2004, during the task observation phase of the DPES Project. Data had been collected from eight volunteers at 2 RAR under similar protocols, with the exception that ballistic vests and helmets had not been worn. As a result, these times were rather faster than for the other trials. The lack of the protective clothing may have also contributed to the higher levels of reliability obtained (ICC = 0.85; Bias = +1%; LOA = 12%). Collectively, these results indicate that the reliability of this test when performed wearing ballistic vests and helmets is near the margin of acceptability for all cohorts of males. IE trainees should undergo thorough familiarisation before being tested. The pooled results for 21 males (trials 1 & 2 at 2 RAR, AFDW and SOI; 6 RAR omitted) were: ICC = 0.84, significant; Bias -5%, significant; LOA = 22%. This was the only test for which the bias in the pooled results was statistically significant (p=0.04).

Predictive Tests

4.3.13 *Loaded Incremental Velocity Run*. This test is scored using the same structure of stages within levels as for a standard multistage shuttle run (beep test). The score is a decimalised level derived from the level and stage reached whilst keeping pace with the externally set incrementally increasing velocity. Whilst ICCs were low for both AFDW and 6 RAR, LOA was within acceptable limits in both cases; bias was acceptable at 6 RAR and marginally unacceptable at AFDW (AFDW: ICC = 0.62; Bias = +8%; LOA = 19%. 6 RAR: ICC = 0.66; Bias = -5%; LOA = 20%). IE trainees at SOI produced better results for ICC and bias, and a similar result for LOA (ICC = 0.89; Bias = -1%; LOA = 19%). These results are regarded as acceptable (the relatively low ICCs at AFDW and 6 RAR were probably due at least in part to homogeneity of performances rather than unreliability per se – see paragraph 2.5.11). It should be noted that the IE trainees performed at a higher level than the other cohorts on this test, and there was no evidence of a learning effect. This is consistent with the fact that this is a non-CATT-based generic test of physical fitness. Females did not undertake this

test. The pooled results for 22 males (trials 1 & 2 at AFDW, 6 RAR and SOI) were: ICC = 0.80, significant; Bias =2%, not significant; LOA = 22%.

4.3.14 *1.82m Wall Climb (standing start)*. This test is administered in conjunction with the 1.82m Wall Climb (running start), and is similarly assessed using both categorical and quantitative measures. Immediately after each of the permitted three attempts to scale a 1.82m wall with a running start, participants attempt to scale the wall from a standing start. The time limit for each attempt is 30 seconds, but if the participant falls back to the ground the attempt is terminated and scored as a fail.[1] For successful attempts, the same quantitative measures were considered as for the Wall Climb (running start), including:

- time taken on the first attempt
- time taken on the second attempt
- time taken on the third attempt
- mean time for all successful attempts
- fastest (minimum) time of three attempts.

4.3.15 All reliability results for this test were very poor. As for the Wall Climb (running start), the fastest time recorded from three attempts was found to be the most reliable (or least unreliable) of all the measures investigated. Results based on this measure are shown in Table 7.

4.3.16 None of the 10 females tested was able to scale the wall from a standing start. The same was true of two of the eight IETs tested. Two of the eight soldiers tested at 6 RAR succeeded in scaling the wall in only one out of six attempts. The fact that this test is beyond the capability or close to the limits of capability of many of the participants reduced the effective sample size for this test. Based on the data from those in the three male cohorts who did scale the wall, in no case did all three reliability measures fall within acceptable limits. It is concluded that the quantitative Wall Climb (standing start) test (with the measure being the fastest time recorded from three attempts) is unreliable and of no value.

[1] Initially, multiple attempts were permitted within each 30 second period ("attempts within attempts"). However, after some initial analysis, it was decided that more reliable time scores would be obtained if multiple attempts were not permitted. This is because someone who succeeds on the first attempt might take say 7 sec. On a subsequent occasion, he/she might miss his/her footing or hit the wall awkwardly, and hence miss at the first attempt, return to the starting point and subsequently succeed on the second jump in say 8 seconds, but with a recorded total elapsed time of 16 seconds – a very different time from the first occasion. All the data were re-analysed using the single attempt protocol.

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

5.1.1 The outcomes of reliability testing are summarised in Table 8. The last column shows the overall assessment of reliability for each test, determined with reference to the indicative target levels (paragraph 2.5.21). Considering the practical constraints on the conditions of testing outlined in section 2.4 above and discussed further below, the levels of reliability of most tests are regarded as either acceptable or provisionally acceptable.

Table 8. Summary of Results of Reliability Testing

Test Name	Infantry soldiers	AFDW airmen	Infantry IETs	Army females	Overall Assessment
Simulation tests					
Tests common to Infantry and ADG					
Pack Lift and Place	×	×	NT	NT	U
Box Lift and Place	NT	NT	✓	NT	PA
Jerry Can Carry	✓	✓	✓	×	A
1.82m Wall Climb (running start) Categorical test: pass/fail	+	+	✓	✓	A
1.82m Wall Climb (running start) Quantitative test: duration	✓	✓	✓ Famil. req.	?	A
Leopard Crawl	✓	✓	✓ Famil. req.	✓ Famil. req.	A
Urban Rushing	?	?	✓ Famil. req	× Famil. req	PA
Section Attack	?	?	✓ Famil. req.	NT[1]	PA
Tasks/Tests for ADG only					
Sustained Patrol (5km)	NA	+	NA	NA	ND
Pursuit (2.4km)	NA	+	NA	NA	ND
Predictive tests					
Tests common to Infantry and ADG					
1.82m Wall Climb (standing start) Categorical test: pass/fail	+	+	✓	–	PA
1.82m Wall Climb (standing start) Quantitative test: duration	×	×	×	–	U
Loaded Incremental Velocity Run	✓	✓	✓	NT	A
Test for Infantry only					
Forced March (10km)	+	NA	+	NT[1]	ND

[1] Females undertook these tests on only one occasion, so no reliability analysis was possible.

Key:

✓ Reliability satisfactory	+ All participants passed test	A Reliability acceptable
× Reliability not satisfactory	– All participants failed test	U Reliability unacceptable
? Reliability marginal or uncertain	NA Not applicable	PA Reliability provisionally acceptable
	NT Not tested	ND No discrimination - no reliability assessment

5.1.2 The best reliabilities, as indicated by low LOAs, were generally obtained with IETs at SOI, suggesting that this cohort in particular tended to put in very consistent efforts. In the CATT-based tests, IETs tended to improve as a group from trial 1 to trial 2. This did not occur in the generic (ie. non CATT-based) Loaded Incremental Velocity Run test. This supports the notion that the general group improvement effect observed in the CATT-based tests had a substantial learning component. This in turn indicates the need for thorough familiarisation with CATT-based tests in the normative testing phase. In the case of exhaustive tests, this requires a separate familiarisation session at least 48 hours prior to testing. This should be taken into consideration when planning the schedule of normative data collection with IETs from both Infantry and AFDW.

5.1.3 Some of the results of the first round of reliability testing at AFDW appeared to be affected by the requirement for ADGs to become familiar with tasks that are performed less regularly by ADG than Infantry. Results from 6 RAR were affected by problems with attitude and compliance with protocols on the part of PTIs, and by the lack of motivation and poor physical condition of the soldiers, many of whom were carrying undisclosed injuries from previous Infantry training activities.

5.1.4 Considering these limitations, it is considered that the results of this study indicate that the reliability of most of the PETs is acceptable or provisionally acceptable. However, the moderate reliability established for a number of these tests indicates that there is the potential for a substantial degree of variation to occur from occasion to occasion in operational use. For this reason it is imperative that adequate familiarisation occurs prior to testing, and that there is adequate opportunity for retesting if a test is failed. It is anticipated that the former will occur since the tests will be used as a basis for training, and that the latter will occur as a matter of course within the competency testing paradigm.

5.1.5 Because of the small number of females recruited for reliability testing, and the fact that not all those recruited were physically capable of safely undertaking the more physically demanding tasks, it was not possible to complete a full program of testing in the limited time available. Furthermore, the extremely small sample size for some tests resulted in greater statistical uncertainty in sample estimates than was the case for the three male cohorts. Of the five tests[1] for which reliability assessments were possible for this cohort, one (Leopard Crawl) was found to be reliable (but requiring thorough familiarisation), and one (Wall Climb with Running Start) was found to be marginally reliable. No females were able to complete a Wall Climb with Standing Start. Reliability was not established for two tests (Jerry Can Carry and Urban Rushing). The data for females exhibited both changes from trial to trial in the performance level of the group as a whole, and large trial to trial variation in individuals. The former is probably indicative of learning effects; the latter may be related to both unfamiliarity with the tasks and how to approach them, and to issues with poorly fitting equipment and apparel, particularly ballistic vests and helmets. It is clear from the reliability testing that the female volunteers were far less familiar with combat arms trade tasks than were the three male cohorts, and that this unfamiliarity was exacerbated by problems with equipment and apparel. It is considered that females would have to undertake an extensive program of familiarisation with these tests in order for reliable measurement to be possible. It is also considered that this is likely to be the case for the other tests for which no reliability testing was undertaken with females. It is of concern to the DPESP research team that our ability to safely collect reliable data concerning female performance on CATT-based tasks in the normative phase of the project is likely to be limited both by a shortage of volunteers with sufficient physical capabilities and by inadequate preparation and equipping of participants.

5.1.6 Notwithstanding the last sentence of the previous paragraph, it should be noted that in the four tests completed by females, the level of performance of the females was significantly lower than that of all the male cohorts. In two of these tests (Jerry Can Carry and Leopard Crawl) there was no overlap at all in the distributions of male and female scores – in each case the best individual female performance was worse than the worst individual male performance. In the other two tests (Wall Climb with Running Start and Section Assault), the best female performances were comparable with the worst male performances. In the fifth

[1] Two other tests, the 10 km Forced March and the Section Assault were also undertaken by females on one occasion, but for safety reasons the second trial of these tests was not undertaken.

test attempted by females (Wall Climb with Standing Start) none completed the test, whereas the great majority of males did so on most attempts. These differences are demonstrated in the results for the Wall Climb with Standing Start (Table 6), and in the mean scores for each cohort on the other four tests undertaken by females (Table 7). In general, in all five tests undertaken by females, there were clear differences between male and female performances, even though the female performances were generally measured with lower reliability than the male performances. However it is not possible to draw any conclusion as to the degree to which these results reflect inherent gender differences in physical capacity. Females in this study were less engaged in physical training on a day-to-day basis than were participants in the three male cohorts, and had had less exposure to the combat trade tasks on which most of the tests were based. As a result, gender differences are confounded with the effects of training and familiarisation. However, it should also be noted that there were two levels of selectiveness operating with the female cohort. Firstly, whilst all participants were volunteers, the element of self selection was much stronger with females, for whom the testing program was unrelated to their current employment category and peer group; and secondly, there was a two-stage fitness hurdle controlling entry into the test program. These mechanisms would presumably have biased the female sample towards better than average female performance.

5.2 Recommendations

5.2.1 The reliability of the following tests (common to Infantry and ADG) is assessed as acceptable for adoption as PETs. It is recommended that data on these tests be collected in the Normative Data Collection phase of the DPES Project:

- Jerry Can Carry
- 1.82m Wall Climb from Running Start (categorical and quantitative: pass/fail and duration)
- 1.82m Wall Climb from Standing Start (categorical test: pass/fail)
- Leopard Crawl
- Loaded Incremental Velocity Run

5.2.2 Whilst the results for the following tests at SOI were encouraging, overall the reliability of these tests could only be assessed as provisionally acceptable. It is recommended that data on these tests be collected in the Normative Data Collection phase of the DPES Project. However, it is also recommended that if any of these tests is adopted by Defence, further reliability testing should be conducted with thoroughly familiarised and motivated participants in good physical condition, in order to confirm the provisional conclusions reached in this report.

Tests common to Infantry and ADG
- Box Lift and Place
- Urban Rushing

Test for Infantry only
- Section Attack after Forced March (10km)

Test for ADG only
- Section Attack after Pursuit (2.4km)

5.2.3 The reliability of the following test could not be assessed because all participants passed the test on both occasions and hence the capacity of the test to discriminate was not established[1]. It is recommended that this test not be further considered in the Normative Data Collection phase of the DPES Project.

- Sustained Patrol (5km)

[1] Whilst the same is true of the 10 km Forced March and the 2.4 km Pursuit, these tests are not performed in isolation, but rather are carried out as pre-fatiguing activities in conjunction with the Section Attack test.

5.2.4 The reliability of the following tests is assessed as unacceptable for adoption as PETs. It is recommended that these tests not be further considered.

- Pack Lift and Place
- Wall Climb from Standing Start (quantitative test: duration)[1].

5.2.5 In order to optimise the reliability of the data to be collected from IE trainees and females in the forthcoming normative phase of the DPES Project, it is recommended that an extensive program of familiarisation with CATT-based tests should be undertaken by these cohorts prior to actual testing. This requirement should be taken into consideration when planning the schedules for normative data collection for IETs from both Infantry and AFDW, and for females.

5.2.6 It is recommended that steps are taken by Defence to ensure that female participants are equipped with well-fitting ballistic vests & helmets during both familiarisation and normative testing.

[1] This does not imply that the Wall Climb from Standing Start should be omitted altogether in the normative data collection phase, but that it should be scored only as a categorical pass/fail test.

REFERENCES

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1: 307-310.

Boldovici, J.A., Bessemer, D.W. & Bolton, A.E. (2001). *The Elements of Training Evaluation*. The U.S. Army Research Institute for the Behavioral and Social Sciences. Accessed from the World Wide Web http://www.hqda.army.mil/ari/pdf/bk2002-01.pdf 21 March 2005.

Cooper, S-M., Baker, J.S., Tong, R.J., Roberts, E., & Hanford, M. (2005). The repeatability and criterion related validity of the 20 m multistage fitness test as a predictor of maximal oxygen uptake in active young men. *Br J Sports Med* 2005, 39: e19. (http://www.bjsportmed.com/cgi/content/full/39/4/e19).

Deuster, P.A., Jones, B.H., & Moore, J. (1997). Patterns and risk factors for exercise-related injuries in women: A military perspective. *Military Medicine,* 10: 649-655.

Engelman, M.E. & Morrow, J.R.,Jnr (1991). Reliability and skinfold correlates for traditional and modified pull-ups in children grades 3-5. *Research Quarterly for Exercise and Sport*, 62: 88-91.

Goldstein, I.L., Zedeck, S., & Schneider, B. (1993). An exploration of job analysis-content validity process. In N. Schmitt & W.C. Borman (Eds.), *Personnel Selection in Organizations*. San Francisco, CA: Jossey-Bass.

Innes, E. & Straker, L. (1999). Reliability of work-related assessments. *Work*, 13: 107-124. Accessed from the World Wide Web http://homepages.wmich.edu/~dhazel/OT481/OT481HomePage/1999Work4Reliability.html 21 March 2005

Jones, B.H., Boyce, M.W., & Knapik, J. (1992). Associations among body composition, physical fitness and injury in men and women Army trainees. In: *Body Composition and Physical Performance*, pp 141-172. Edited by Marriott, B.M, Gunstrup-Scott, J. Washington DC, National Academic Press.

Knapik, J., Cuthie, J, Canham M., Hewitson, W., Laurin, M.J., Nee, M.A., Hoedebecke, E., Hauret, K., Carrol, D., & Jones, B. (1997). *Epidemiological consultation No. 29-HE-7513-98: Injury incidence, injury risk factors, and physical fitness of U.S. army basic trainees at Ft. Jackson, South Carolina.* (Technical Report). Aberdeen Proving Ground, MD: US Army Centre for Health Promotion and Preventive Medicine, Directorate of Epidemiology and Disease Surveillance.

Mathiowetz, V. (2002). Comparison of Rolyan and Jamar dynamometers for measuring grip strength. *Occupational Therapy International*, 9(3): 201–209.

McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psyc*hological *Methods,* 1(1): 30-46.

Ministry of Defence. (Undated). *Fit to Fight Pamphlet 2: Test Protocols and Administrative Instructions.*

Pandolf, K.B., Givoni, B. & Goldman, R.F. (1977). Predicting energy expenditure with loads while standing or walking very slowly. *Journal of Applied Physiology,* 43: 577-581.

Pandorf, C.E., Nindl, B.C., Montain, S.J., Castellani, J.W., Frykman, P.N., Leone, C.D. and Harman, E.A. (2003). Reliability assessment of two militarily relevant occupational physical performance tests. *Canadian Journal of Applied Physiology*, 28(1): 27-37.

Pate, R.R., Burgess, M.L., Woods, J.A., Ross, J.G., & Baumgartner, T. (1993). Validity of field tests of upper body muscular strength. *Research Quarterly for Exercise and Sport*, 64: 17-24.

Portney, L.G. & Watkins, M.P. (1993). *Foundations of Clinical Research: Applications to Practice.* East Norwalk, CT, Appleton & Lange.

Rayson, M., Wilkinson, D., Carter, J., Richmond, V. & Blacker, S. (2004). *The Reliability of the RAF's Proposed Representative Service Tasks*. Bristol, UK, Optimal Performance Limited.

Rosen, G.A., (1978). The problem and utility of work sample reliability data. *Vocational Evaluation & Work Adjustment Bulletin,* 11(3): 45-50.

Sparling, P.B., Millard-Stafford, M., & Snow, T.K. (1997). Development of a cadence curl-up test for college students. *Research Quarterly for Exercise and Sport,* 68: *309-316.*

Virginia Department of Education. (1999). *Standards of Learning (SOL) Tests Validity and Reliability Information Spring 1998 Administration*. Virginia Department of Education Division of Assessment and Reporting. Accessed from the World Wide Web http://www.pen.k12.va.us/VDOE/Assessment/validity.PDF 21 March 2005.

Women in the Armed Forces Steering Group. (2002). *Women in the Armed Forces*. Report of the Women in the Armed Forces Steering Group, UK Ministry of Defence. Accessed from the World Wide Web: http://www.mod.uk/linked_files/ewaf_full_report.pdf August 19, 2004.

Young, W., Macdonald, C., Heggen, T. & Fitzpatrick, J. (1997). An evaluation of the specificity, validity and reliability of jumping tests. *Journal of Sports Medicine and Physical Fitness*, 37: 240-5.

ANNEXES

Annex 1

RELIABILITY TESTING ADG

TIME	MON	TUE	WED	THU	FRI
0700-0900	ADMIN/FAMIL/13 - ALL	8A	5B	8A	5B
0930-1030		7B	6A	7B	6A
1100-1200	3A/4A	3A/4A	1B	RESERVE	1B
1330-1530	2B	2B	RESERVE	1B	RESERVE

Legend

ID	ACTIVITY	ID	ACTIVITY
1	Pack Lift and place onto a UNIMOG	**6**	5 km Sustained Patrol
2	Jerry Can Lift and Carry	**7**	Loaded Incremental Velocity Run
3	1.8m Wall Climb	**8**	ADG Pursuit and Assault
4	Leopard Crawl	**9-12**	Reserved
5	Urban Rushing	**13**	Anthropometry Measurements

A/B refers to sections ie 3B means section B undertaking activity 3

Annex 2

RELIABILITY TESTING 6 RAR

TIME	MON	TUE	WED	THU	FRI
0730-0900	ADMIN/FAMIL/13 - ALL	6B	5A	6B	5A
0930 - 1030			RESERVE		1B
1100-1200	2B	7A	2B	7A	RESERVE
1330-1530	3A/4A	1B	3A/4A	1B	RESERVE

Legend

ID	ACTIVITY	ID	ACTIVITY
1	Pack Lift and Place onto a UNIMOG	**6**	10km March and Assault
2	Jerry Can Lift and Carry	**7**	Loaded Incremental Velocity Run
3	1.8m Wall Climb	**8-12**	Reserved
4	Leopard Crawl	**13**	Anthropometry Measurements
5	Urban Rushing		

A/B refers to sections ie 3B means section B undertaking activity 3

Annex 3

RELIABILITY TESTING WOMEN

TIME	MON	TUE	WED	THU	FRI
0730-0900	ADMIN/FAMIL/13 - ALL	6B	5A	6B	5A
0930 - 1030			RESERVE		1B
1100-1200	2B	7A	2B	7A	RESERVE
1330-1530	3A/4A	1B	3A/4A	1B	RESERVE

Legend

ID	ACTIVITY	ID	ACTIVITY
1	Pack Lift and place onto a UNIMOG	**6**	10km March and Assault
2	Jerry Can Lift and Carry	**7**	Loaded Incremental Velocity Run
3	1.8m Wall Climb	**8-12**	Reserved
4	Leopard Crawl	**13**	Anthropometry Measurements
5	Urban Rushing		

A/B refers to sections ie 3B means section B undertaking activity 3

Annex 4

RELIABILITY TESTING SCHOOL OF INFANTRY

TIME	MON	TUE	WED	THU	FRI
0730-0900	ADMIN/FAMIL/13 - ALL (*1)	6B	5A	6B	5A
0930 - 1030			3B/1B (O830 start)		3B/1B (O830 start)
1100-1200	2A (*2)	7A	4A	7A	2A
1330-1530	3B/1B	4A	2A	4A	RESERVE

Legend

ID	ACTIVITY	ID	ACTIVITY
1	Box Lift and Place onto a UNIMOG – hard stand	**6**	7 km March and Assault - oval
2	Jerry Can Lift and Carry - oval	**7**	Loaded Incremental Velocity Run - oval
3	1.8m Wall Climb - obstacle course	**8-12**	Reserved
4	Leopard Crawl - oval	**13**	Anthropometry Measurements - classroom
5	Urban Rushing - oval		

*1 Personnel to be in PT Kit. They are to bring webbing, packs etc to enable weigh in for all activities that they are involved in.

*2 A/B refers to sections ie 3B means section B undertaking activity 3

For all activities, less 6 and 7, all up weight including webbing, helmet, ballistic vest and including personal weapon, is 21.6 kg.

For activity 6, all up weight, including pack, webbing and personal weapon is to be 25 kg

For activity 7, all up weight including webbing and personal weapon is to be 12.5 kg (if this weight is unable to be fitted into webbing then the pack or day pack may be used)

Activity	Equipment Requirements to be supplied by SOI
1	Unimog, witches hats, seven 5kg weights
2	witches hats, line marking
3	witches hats
4	witches hats
5	witches hats, line marking
6	witches hats, 50 star pickets, tape, line marking
7	witches hats, line marking

Annex 5. Information Sheet

DEFENCE PHYSICAL EMPLOYMENT STANDARDS PROJECT

TEST DEVELOPMENT FIELD TRIALS – RELIABILITY STUDY

INFORMATION SHEET

The purpose of this letter is to describe the "Defence Physical Employment Standards Project: Test Development Field Trials: Reliability Study" and to invite you to participate in the study.

Brief description of the trials

There is no doubt that you have been trained to do a variety of highly skilled tasks that require a range of physical as well as technical skills. It is also well recognised that it is important to make sure that you are physically capable of completing the tasks that you will be required to do.

These trials are part of a larger project that aims to establish a number of clearly justified physical employment standards or fitness tests for the combat arms trades. The tests should be seen as being separate from the Basic Fitness Assessment (BFA) and Combat Fitness Assessment (CFA) tests that you currently undertake. In general, the purpose of the tests that will be developed from the larger project is to ensure that soldiers allocated to the various parts of the Australian Defence Force (ADF) are physically able to do their job in a safe and effective way. In the first instance, the subjects of the study are Infantry and Airfield Defence Guards (ADG).

The purpose of the trials is to determine reliability (ability to obtain the same result on different occasions) of 9 tests for Infantry soldiers, Infantry initial employment trainees (IETs) and female soldiers; and 10 tests for ADG airmen. The tasks and tests have been chosen by a group of experienced soldiers and airmen who believe that the tests represent the most physically demanding parts of the job of being an Infantry soldier or Airfield Defence Guard. The tests and test procedures are listed on sheets attached to this information sheet.

Your part in the study

I would like to invite you to participate in this study. It is important for you to note that your involvement in the study is entirely voluntary and if you chose not to participate there will be no detriment to your career or future health care. Finally, if you chose to participate and later

change your mind and wish to withdraw, you may do so without any detriment to your career or future health care. Non-participants will be assigned to alternative duties within the regular framework of work & training cycles.

The study has been scheduled into your Unit's work plan and will take approximately one week to complete in the period from November 2004 to March 2005. If necessary, you will be given an opportunity to rehearse a test in order to familiarise yourself with the procedures and then you will perform the task as part of a 10 person section. The various tests will be organised to make sure that you remain as fresh as possible throughout the study.

Up to 40 Infantry soldiers, 40 Infantry IETs, 40 female soldiers and 40 Airfield Defence Guards will be recruited to participate in this study.

The soldiers/guards will be experienced and well trained, and will not be asked to do anything that they would not normally do as part of their job. Where required on the advice of ADF specialists, tests to be performed by IETs and female soldiers have been adjusted to take account of the lower physical capacities of these groups.

You are asked to do your best in each test. The tests include some fitness tests that are very similar to the Basic Fitness Assessment that will be conducted on all participants at the start of the study. I would also like to point out that in order to try to control for the level of fatigue you experience during the study, your participation in exercise/training outside of the study will be controlled with reasonable limits as will your general diet for the duration of the study.

A number of measurements will be made and records taken while you perform each of the tests. These include measuring heart rate from a heart rate meter strapped around your chest, video taping each of the tests and asking you to complete a simple 'pencil and paper' test to measure the effect of the task on your level of tiredness (rating of perceived exertion).

Risks of participating

It is important to point out to you that there will be a number of risks associated with participation in this study. However, as you would expect, a range of safeguards have been put in place to make sure that these risks will be minimised.

The first risk is that you feel that you are being coerced or forced to participate in this study. In order to minimise the potential for coercion, recruitment of volunteers will be conducted by

a person who is not in your direct chain of command. As mentioned above, you will also be formally notified of your freedom to withdraw at any time should you change your mind about participating in the study. Non-participants will be assigned to alternative duties within the regular framework of work & training cycles.

Secondly, participation in the study will expose you to a degree of injury risk. A number of safeguards have been put into place to minimise this risk.

1. You will not be able to participate if you are carrying an injury or have an illness that may be made worse as a result of your involvement in the study. You will be asked to disclose your injury and illness status to a civilian researcher and the information that you report will be kept in confidence within the University of Ballarat research team.
2. If you are an experienced soldier or guard in a combat unit, you will not be asked to undertake anything that is different from that which you do in your normal job. If you are an IET trainee or a woman, where appropriate tasks have been modified to reflect your level of job training and physical capability.
3. All tasks will be monitored by an experienced Warrant Officer to make sure that they are done in the safest possible way.
4. The area in which you will perform the tasks will be checked to make sure that there are no unacceptable physical hazards present.
5. Heat injury: Soldiers and airmen performing heavy work tasks experience an increase in body temperature. In some circumstances an increase in body temperature may result in a soldier or airman/airwoman being placed at an unacceptable risk of incurring a heat injury and in rare instances death. A number of preventive and treatment strategies will be implemented to ensure that the risk of heat injury and adverse consequences is minimised during the activities involved in the DPES Project. These strategies will involve the following:
   - Fitness: Only physically fit soldiers and airmen/airwomen will participate in the study. All male soldiers will have satisfactorily completed the combat arms CFA, airmen the BET, trainees the recruit training exit test, and female participants the relevant BFA, CFA or PFT. Women undertaking endurance tests will have attained a high level on the multistage shuttle run test.
   - Heat Acclimatization: It is well accepted that participants acclimatized to the heat will be more able to tolerate work in hot conditions than those who are unacclimatized. In the present study, soldiers and airmen currently work and live in the climate where they are being tested. Trainees will have recently moved from Kapooka to Singleton and from Edinburgh to Amberley. In neither case is the climatic difference in May-July regarded as presenting a significant acclimatization issue. Women participants will be

resident in the testing location but will not be as accustomed to the exercise/work regimen as the soldiers/airmen. Women will be provided with a recommended program of preparatory exercise to be undertaken in the period between recruitment and the commencement of testing.

- Minimizing environmental heat load: Exercising in a cool environment obviously will result in a reduced environmental heat load and a increased ability for you to lose the heat produced during exercise. The project has been designed to take advantage of the cooler temperatures experienced in northern Australia from April to August. In addition, the tasks likely to induce the greatest heat load (20 km and 10 km forced marches) will be conducted in the morning with the first group of troops 'stepping off' at 0500 hr.
- Core temperature limits: Core body temperature may be measured and for tests in which there is a risk of unacceptably raised core temperatures. This may include some or all of the endurance tests (7 km forced march, section attack, urban rushing, loaded incremental velocity run).
- Hydration: You will be advised to ingest 500 ml of fluid up to 2 hr prior to exercise and up to 1.2 litres per hour of cool fluid during exercise.
- Heart rate limits: A heart rate limit of 90% of age predicted maximal heart rate (220 beats/min – age in years) will be applied.
- Signs of heat intolerance: You will be monitored by the research staff and the attending ADF medical staff for signs of heat intolerance. These include the presence of red, hot, dry skin; throbbing headache, dizziness, nausea and confusion.
- Availability of first aid/medical support. An ADF first aid team will be present and able to provide appropriate support during the 20 km and 10 km forced marches and section attack (Infantry) and during the sustained patrol, pursuit and section attack activities (ADG). In addition, the activity will be conducted on bases where medical support is immediately available should it be required.
- Recent medical history. You will be required to complete a 'Confidential Health Status Form' which includes questions regarding factors that may predispose a participant to heat stress such as medications, viral illness (fever), cardiac conditions and diabetes If you report any of these conditions/factors you will be unable to participate in the project.

If you do experience any type of injury, you will be given the first aid or medical treatment necessary by qualified personnel.

This study requires that very little equipment will actually be fixed to your body. The only piece of equipment that will be fixed to your body is a heart rate meter. This consists of a strap placed around your chest (heart rate transmitter) and a receiver that looks like a wrist watch. Both pieces of equipment use watch batteries for power and have been passed as safe for use by humans. Core temperature may be measured using pills which you swallow and which pass through your digestive tract. Use of these pills is not advised if you have particular characteristics or medical conditions. These are listed on the health status form which you will asked to complete before participating in this project. If you have any of these conditions you will not be allowed to participate in tests involving core temperature monitoring. Other data will be collected from pencil and paper tests such as recording how tired you feel. Finally, a video record will be made of you when you undertake each of the work tasks.

Statement of Privacy

There is a separate risk associated with protecting your privacy. There is a risk that the data collected may be used inappropriately within Defence or within the wider community. Examples of this may include using a photo of you without your permission or quoting your individual results in a Defence report. These risks will be reduced by the following:

1. You will be given a code number specific to this study and all data will be 'de-identified' whereby your name will be removed from any sets of records that are used for analysis and reported on to Defence or distributed in the wider community.
2. The information that links your name to your code will be held in confidence by the civilian Principal Researcher.
3. Only group data summaries will be used in any reports
4. Any videos or pictures that are included in the reports will be 'de-identified' by blurring your face or the Civilian Chief Investigator will seek your written permission to use the original image if this is considered desirable.
5. All original data will be kept under lock and key at the University of Ballarat for a period of at least five years.
6. Secure information disposal methods will be used such as document shredding.
7. The data will only be used for the purposes outlined above without your express permission.

On duty

All members of the Australian Defence Force who volunteer for this study will be considered to be on duty when participating in the study.

Names of Investigators

Principal Investigators:

Professor Warren Payne
School of Human Movement and Sport Sciences
University of Ballarat
PO Box 663
Ballarat, 3350
Telephone: (03) 5327 9693
Email: w.payne@ballarat.edu.au

Dr. Jack Harvey
School of Information Technology and Mathematical Sciences
University of Ballarat
PO Box 663
Ballarat, 3350
Telephone: (03) 5327 9273
Email: j.harvey@ballarat.edu.au

Should you have any complaints or concerns about the manner in which the project is conducted, please do not hesitate to contact the researchers listed above in person or you may prefer to contact the Australian Defence Human Research Ethics Committee or the University of Ballarat Human Research Ethics Committee at the following addresses:

Executive Secretary
Australian Defence Human Research Ethics Committee
CP2-7-66
Department of Defence
CANBERRA ACT 2600
Telephone: (02) 6266 3837
Facsimile: (02) 6266 4982
Email: ADHREC@defence.gov.au

Executive Officer
University of Ballarat Human Research Ethics Committee
Office of Research
University of Ballarat
PO Box 663
Ballarat, 3350
Telephone: (03) 5327 9765
Facsimile: (03) 5327 9602
Email: k.bernard@ballarat.edu.au

Annex 6. Consent Form

DEFENCE PHYSICAL EMPLOYMENT STANDARDS PROJECT

TEST DEVELOPMENT FIELD TRIALS – RELIABILITY STUDY

INFORMED CONSENT FORM

I,…… give my consent to participate in the project mentioned in the subject information sheet on the following basis:

I have had explained to me the aims of this research project, how it will be conducted and my role in it.

I understand that I am participating in this project in a voluntary capacity and can withdraw at any time without penalty or detriment to my career or future health care.

I understand that, as an ADF member, I will be considered to be 'on duty' during participation in the study.

I understand the risks involved as described in the subject information sheet.

I am co-operating in this project on condition that:

- The information I provide will be kept confidential.
- The information will be used only for this project.
- The research results will be made available to me at my request and any published reports of this study will preserve my anonymity.

I have been given a copy of the information/consent sheet, signed by me and by the principal researcher, Prof. Warren Payne, to keep.

I have also been given a copy of ADHREC's *Guidelines for Volunteers*.

Video clips and still shots may be used for reports and presentations. If clips or shots are used you may be identifiable. Please tick one of the following options:

☐ I give permission to use video clips or still shots that identify me.

☐ I give permission to use video clips or still shots where my face is pixellated (thus de-identifying me).

☐ I DO NOT give permission to use video clips or still shots of me.

……	……
Participant's signature	Principal Researcher's signature
……	……
Printed name	Printed name
……	……
Date	Date

Annex 7. Health Status Form

DEFENCE PHYSICAL EMPLOYMENT STANDARDS PROJECT

TEST DEVELOPMENT FIELD TRIALS – RELIABILITY STUDY

CONFIDENTIAL HEALTH STATUS FORM

Name: _____ Age:_____

Weight: _____ kg Height: _____ cm

Please circle the best answer to the questions below and provide details where requested

1. Do you smoke? No Yes

2. Have you smoked in the past? No Yes – please give details

_____

3. Are you currently overweight or obese? No Yes Don't know

4. Do you currently have:

high blood pressure	No	Yes	Don't know
diabetes	No	Yes	Don't know
asthma	No	Yes	Don't know
haemophilia	No	Yes	Don't know

5. Do you or anyone in your family have a history of cardiovascular disease (heart attack, chest pain, stroke)?
No Yes Don't know

If YES, please give details: _____

6. Are you currently on any medication?
No Yes

If YES, please give details:_____

7. Do you have a current or ongoing muscle, joint or bone injury that may prevent you from participating in and completing this study?
No Yes

If YES, please give details:_____

8. Do you have any other medical complaint or any other reason which you think may prevent you from participating in and completing this study?
No Yes

If YES, please give details:_____

PLEASE TURN OVER TO COMPLETE THE FORM

9. Do you know or have you ever been informed that you have any of the following which would prevent you from using core temperature pills?

Body weight less than 36 kg	No	Yes
Obstructive disease of the gastro-intestinal tract	No	Yes
Gag reflex impairment	No	Yes
Have undergone gastro-intestinal surgery	No	Yes
Felinization (transverse folds) of the esophagal mucosa	No	Yes
Hypomotility disorder of the gastro-intestinal tract	No	Yes
A cardiac pacemaker or other implanted electromedical device.	No	Yes
Will undergo magnetic resonance imaging within the next three weeks	No	Yes

Signed: _____ Date: _____

Print Name: _____

Annex 8. Test Procedures

DEFENCE PHYSICAL EMPLOYMENT STANDARDS PROJECT

TEST DEVELOPMENT FIELD TRIALS – RELIABILITY STUDY

TEST PROCEDURES

The purpose of this Reliability Study is to assess the reliability of a series of potential tests of physical job performance capability that have been designed by the research team in consultation with Infantry and Airfield Defence Guard (ADG) personnel. These tests are either simulations of actual Infantry and/or ADG job tasks or tests that have been designed to predict your performance on key Infantry and/or ADG job tasks. The key job tasks that these tests have been designed to simulate or predict are called criterion tasks.

CRITERION TASKS AND THEIR RELATED TESTS

1. Loading a UNIMOG Truck (Infantry and ADG): Box Lift and Place Test

Loading a UNIMOG truck is a standard and commonly practiced trade task that requires strength and strength endurance fitness qualities. This task also encompasses a number of aspects of the company level replenishment trade task. The test designed to simulate the loading of the UNIMOG is the Box Lift and Place Test. The participants will be required to lift and place a box containing a graduated sequence of weights onto a platform at a height of 1.51 m $\pm$ 2.1 cm above the ground - the same height as the the back of a UNIMOG truck. The box weight will be progressively increased from 15 kg to a maximum of 45 kg in increments of 5 kg. The participants are to wear patrol order, flak jacket and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg). In order to reduce the risk of departure from good (ie. safe) lifting and carrying technique, the test will not be a test of speed. Rather, participants will be required to work at an externally set pace, loading a box every 30 seconds. The measure of the test is the weight of the heaviest box lifted while working at the required speed and while displaying good (ie. safe) lifting technique as determined by the test administrator.

2. Jerry Can Carry (Infantry and ADG): Jerry Can Lift and Carry Test

The jerry can carry is the second commonly practiced activity undertaken by the Infantry soldier/ADG airman. The test designed to simulate this task is the Jerry Can Lift and Carry. It also simulates aspects of a stretcher carry and company level replenishment. The test involves greater levels of strength endurance than the Box Lift and Place Test and involves a different range of movements. All participants will be required to carry two jerry cans, each filled with water and weighing a total of 22 kg, as far as they are able. The test will involve walking with the jerry cans for distances of 100 m at a velocity of 5 $k.hr^{-1}$. An unloaded walk of 100 m at 5 $km.hr^{-1}$ (1.2 min) will be interspersed between each carry to simulate returning to pick up the second set of two jerry cans. The participants are to

wear patrol order, flak jacket and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg). The measure of the test is the distance over which the jerry cans are carried at the required pace and while displaying good (ie. safe) lifting and carrying technique as determined by the test administrator.

3. Climbing a 1.8 m (6') wall (Infantry and ADG): 1.8 m wall climb test

The 1.8 m (6') wall climb with a running approach is a commonly undertaken trade task. It requires speed, leg power and upper body strength. The task also simulates aspects of jumping and stair climbing. A standing 1.8 m wall climb has also been incorporated in this test. It is considered to be more reliant upon leg power and upper body strength than the running 1.8 m wall climb, and as such it is believed to be a surrogate for the 2.4 m wall climb, the grappling hook climb and the 3.6 m wall climb, which have been rejected as infeasible tasks on which to base simulation tests. The test will be administered as follows. All participants will be allowed three attempts to scale the wall with a running start and then scale the wall in the reverse direction from a standing start. The participants are to wear patrol order, flak jacket and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg). The test will be scored as the times taken to scale the wall with running and standing starts.

4. Crawling through a tunnel (Infantry and ADG): Leopard crawl test

Crawling in a prone position (known as leopard crawling) is a task commonly undertaken by infantry soldiers and ADG airmen. The criterion task is crawling through a 50 m tunnel while wearing patrol order, flak jacket and helmet (17.7 kg) and carrying a standard weapon (Steyr, 3.9 kg). The simulation test will involve the participants crawling using the leopard crawling technique for 25 m as fast as possible on a grassed surface while wearing identical clothing and equipment to that worn in the criterion task. The test will be scored as the time taken to complete the 25 m crawl using a technique considered acceptable to the test administrator.

5. Moving through urban terrain (Infantry and ADG): Urban rushing test

Moving through a high threat urban environment requires the infantry soldier or ADG airman to use a technique that involves a series of sprint-recovery intervals that generally involve a sprint over 22 m and a recovery in a crouched position for seven seconds. This process may be repeated many times but it is considered that the criterion distance will be 220 m. This criterion task is referred to as 'urban rushing' and involves a combination of strength, strength-endurance and endurance. The simulation test will involve all participants performing externally paced 20 m sprints (at a constant velocity to be determined, as opposed to the increasing velocities of a shuttle run) interspersed with a standardised recovery period of seven seconds undertaken in a crouched position. All participants are to wear patrol order, flak jacket and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg). The test terminates when a participant fails to complete two consecutive sprints at the required velocity. The test is scored as the number of sprints completed.

6a. Forced march (Infantry): Infantry Forced march test

The criterion for the infantry Forced March trade task is considered to be comprised of a 20 km march undertaken while each soldier carries a total weight of 45 kg (marching order plus standard weapon – Steyr) and completed in a time of 3 hr 50 min over flat terrain. Body armour (flak jacket and helmet) will not be worn. The predictive test will consist of a march over a distance of 10 km (flat terrain) at an average velocity of 5 km hr^{-1}. A work:rest regimen of 50 min march at 6 km hr^{-1}, 10 min rest will be employed. The marching velocity will be checked every 2.5 km and the soldier will be allocated time demerits should they miss a given time goal. The test score is the number of demerits accumulated by a soldier. The test administrator will remove any participant they consider is at risk of, or who sustains, an injury. The weight carried by trained soldiers will be 45 kg and by IET soldiers will be 25 kg while the women will carry 20 kg.

7a. Section attack (Infantry): Infantry section attack test

The Section Attack is considered to be a key trade task for Infantry as it is for ADGs. It is undertaken by the Infantry using techniques that are different to those used by the ADGs. The Section Attack is a highly tactical operation and is executed in a range of terrain and environmental conditions. In a generalised sense, the task, as completed by the ADGs, involves covering approximately 100 m using two consecutive sub-tasks: 1. fire and movement and 2. fight through. In a generalised sense, the fire and movement activity is conducted over 80 m and involves a repetitive routine that incorporates running in a straight line for 3-5 m, falling to the ground, crawling (leopard crawl) for 3-5 m at an angle to the run, pausing to shoot and await the remainder of the section to move into position. The pause is approximately equal in time to the time of the initial movement period plus 3-5 seconds. The movement period is a maximum of approximately three seconds. The fight through is a leopard crawl for 20 m undertaken at a rapid pace. Whilst the Section Attack is undertaken in a variable manner depending on the tactical needs of the situation, it has been agreed that it will be simulated by performing a standardised routine on a grassed oval. This test will be comprised of a series of 14 repeated 5 m runs and 3 m leopard crawls. The test will be undertaken with the participants wearing patrol order, flak jacket and helmet and carrying a standard weapon (Steyr) throughout the test (21.6 kg). The test score will be the time taken to complete the Section Attack course.

The Forced march and the Section Attack will be integrated into a single activity for the purposes of test administration. The Forced march will be considered a 'pre-fatiguing' activity. The soldiers will be permitted a 10 min recovery period following the completion of the march prior to undertaking the Section Attack test. Although performance on the march will be scored, the key measure will be the time taken to complete the Section Attack.

6b. Pursuit (ADG): Pursuit test

The Pursuit is a criterion task unique to the ADGs and involves a period of running of up to 3 km to pursue an enemy that has been detected in the vicinity of the airfield. The criterion task involves performing the task at a velocity of 9 km hr^{-1} while wearing patrol order, flak jacket and helmet and carrying one's allocated weapon. The task has been standardised to establish a simulation test by undertaking the task as described above for a distance of 2.4 km over flat terrain while carrying a

standard weapon (Steyr). The pursuit velocity will be checked every 500 m and the airmen will be allocated time demerits should they miss a given time goal. The test score is the number of demerits accumulated by an airman. The test administrator will remove any participant they consider is at risk of, or who sustains, an injury.

7b. Section attack (ADG): ADG section attack test

The Section Attack is considered to be a key trade task for ADGs. It is undertaken by the ADGs using techniques that are different to those used by the Infantry. The Section Attack is a highly tactical operation and is executed in a range of terrain and environmental conditions. In a generalised sense, the task, as completed by the ADGs, involves covering approximately 100 m using two consecutive sub-tasks: 1. fire and movement and 2. fight through. In a generalised sense, the fire and movement activity is conducted over 80 m and involves a repetitive routine that incorporates running in a straight line for 3-5 m, falling to the ground, completing a 360^{o} roll (on longitudinal axis), crawling (leopard crawl) for 3-5 m at an angle to the run, pausing to shoot and await the remainder of the section to move into position. The pause is approximately equal in time to the time of the initial movement period plus 3-5 seconds. The movement period is approximately three seconds. The fight through is a leopard crawl for 20 m undertaken at a rapid pace. Whilst the Section Attack is undertaken in a variable manner depending on the tactical needs of the situation, it has been agreed that it will be simulated by performing a standardised routine on a grassed oval. This test will be comprised of a series of 14 repeated 5 m runs and 3 m leopard crawls. The test will be undertaken with the participants wearing patrol order, flak jacket and helmet and carrying a standard weapon (Steyr) throughout the test (21.6 kg). The test score will be the time taken to complete the assault course.

The Pursuit and the Section Attack will be integrated into a single activity for the purposes of test administration. The Pursuit will be considered a 'pre-fatiguing' activity. The airmen will be permitted a 10 min recovery period following the completion of the march prior to undertaking the Section Attack test. Although performance on the march will be scored, the key measure will be the time taken to complete the Section Attack.

8. Sustained patrol (ADG): ADG sustained patrol test

The Sustained Patrol criterion task undertaken by the ADG airmen takes place over a distance of 5 km (flat terrain) at a velocity of 6 $km.hr^{-1}$ while wearing full battle order (heavy). The total weight carried is 45 kg including a standard weapon (Steyr). The simulation test will involve airmen completing the above regimen. The marching velocity will be checked every kilometre and the airmen will be allocated time demerits should they miss a given time goal. The test score is the number of demerits accumulated by an airman. The test administrator will remove any participant they consider is at risk of, or who sustains, an injury.

PREDICTIVE TESTS

Five further tests will be undertaken to determine their predictive capacity for performance on the tests based on trade tasks. These tests, which apply to both Infantry and ADG, are:

- Maximum heaves.
- Jump and reach.
- Situps.
- Multistage shuttle run.
- Loaded incremental velocity run.

The first four of these tests are standardised tests in common use. The loaded incremental velocity run has been developed from the multistage shuttle run test. The shuttle run is a standard test commonly used by the ADF. The test involves running up and down a 20 m track. The velocity of running is increased periodically until the participant is unable to maintain the required running velocity. The score is the number of runs (shuttles) completed. This score can be converted to an estimate of relative or absolute aerobic capacity. The unloaded shuttle run has been shown to be a good predictor of the likelihood of injury among Army recruits. However, it has been suggested that unloaded tests of endurance capacity have a body-size bias and unfairly discriminate against larger individuals, and that performance on a loaded test would remove the body-size bias. Such a loaded test could be externally paced (eg. shuttle run) or self-paced (eg. 3.2 km run). Reliability of self-paced tests is limited by their reliance upon the participants' pacing ability and have largely been replaced by externally paced tests such as the shuttle run. However, a loaded shuttle run would subject participants to high risk of injury due to the requirement for rapid turns under load. Therefore, a loaded externally paced circular run is being developed in the current study. This run will be similar to the shuttle run in that each distance segment will be externally paced at incrementally increasing velocities, but it will require no stopping or turning. Pilot testing will include the development of a feasible protocol for administering this test. The loads are 22.5 kg for trained soldiers and airmen, 12.5 kg for soldiers undertaking IET and 10 kg for women. The data obtained from this test will be used to determine whether performance on the forced march and the subsequent section attack test can be predicted from performance on the loaded run.

KINANTHROPOMETRY

The kinanthropometric procedures adopted in this project are those of the International Society for the Advancement of Kinanthropometry (ISAK). The ISAK Restricted Profile will be used. This includes eight skinfold measurements (triceps, subscapular, biceps, iliac crest, supraspinale, abdominal, front thigh, medial calf), five girths (arm - relaxed, arm - flexed and tensed, waist, hip, calf – maximum) and two bone breadths (humerus, femur).

Annex 9. Test Protocols

DEFENCE PHYSICAL EMPLOYMENT STANDARDS PROJECT

TEST DEVELOPMENT FIELD TRIALS – RELIABILITY STUDY

Test Protocols: Trained Infantry & ADG

1. Box Lift and Place onto a UNIMOG

Task requirement

1. The task requires you to lift and place the 'ammo box' of varying weights on to the back of the UNIMOG located 1.51 ± 2.1 cm above the ground.
2. You are to wear patrol order, ballistic vest and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg).

Purpose

To simulate:

1. Loading and unloading of a UNIMOG
2. Company/squadron level replenishment task

Equipment requirements ADF	Equipment requirements UB
1. Unimog 2. 2 witches hats	1. Recording sheets 2. 3m tape measure 3. Stop watch 4. ammo box with weights 5. Marking paint 6. scales and boards × 2

Specifications

1. You will be required to lift an ammo box weighing 15kg as a warm up/practice. You will have three attempts to complete this satisfactorily before moving onto the task.
2. You will be asked to move from the witches hat (2 metres from truck) to pick up the ammo box and place it on the truck, move back to the cones then wait until the box has been placed back on the ground. You have 30 seconds rest until moving up to the next level. Eg) 20, 25, 30, 35, 40, 45.
3. It is important that you use a two step lift and place. The first step is lifting the ammo box to the waist, take a step forward and place it on the truck. If you deviate from this it is considered poor form and a warning will be given. If you hit the rear gate with the pack then that is a warning as well. Two warnings means that you will, for your own safety, be asked to stop.

Measures

The measure of the test is the maximum weight of the ammo box your were able to successfully lift and place on the back of the UNIMOG while displaying good (i.e. safe) lifting and carrying technique as determined by the test administrator.

Discontinuation criteria

Participants discontinue the task if you:

1. are injured;
2. display incorrect lifting technique (two warnings);
3. have reached a significant personal level of fatigue;
4. feel you can not safely make the lift at the current weight;
5. are not capable of making the next heaviest weight.

Number of participants that can be tested simultaneously

2

Personnel duties

1. 1 PTI to take a warm-up.
2. 2 researchers will be required to administer, coach and determine successful technique requirements. Moreover, they will be required to measure, set up boxes with weights, record weight and video tape. Recording environmental temperatures at start and finish.

2. Jerry Can Lift and Carry

Task requirement

1. The task requires you to pick up the jerry cans and carry them 50 metres then place them back down for five seconds.
2. Then walk again with jerry cans for 50 metres and place them down.
3. You will then be required to walk 50 metres without jerry cans
4. Repeat until failure or until the jerry cans have been carried 600 metres
5. You will be required to wear patrol order, ballistic vest and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg).

Purpose

1. The purpose of this task is to simulate aspects of a stretcher carry and company/squadron level replenishment.

Equipment requirements ADF

1. 10 × 20 litre jerry cans full of water (22 kg each)
2. Oval
3. 10 witches hats
4. 1 × 100 metre long × 10 metre wide (flat and mown)

Equipment requirements UB

1. Marking paint
2. Stop watches
3. Tape measure (30m)
4. Marking paint
5. CD that works in CD player and batteries

Specifications

1. The test will involve you walking with the jerry cans for 2×50 m at a velocity of 5 $km.hr^{-1}$.
2. A whistle or a beep from the CD player will make sure you are maintaining correct pace. You will hear a sound when you should be at the next witches hat which is 50 metres away and you will have 40 seconds to achieve this in. At the cone you will place the jerry can down and rest for 5 seconds.
3. You will then need to pick up the jerry can again and walk another 50 metres in 40 seconds and place the jerry can down.
4. You will then need to complete an unloaded walk of 50 metres in 30 seconds. Once back to the jerry cans you have 5 seconds before completing the task again.
5. You are to continue with the task until volitional fatigue or until the jerry cans have been carried 600 metres.

Measures

The measure of the test is the distance and time over which the jerry cans are carried at the required pace and while displaying good (ie. safe) lifting and carrying technique as determined by the test administrator.

Discontinuation criteria

You will be ask to discontinue with the task if you:

1. cannot complete the task in the required times (miss two cones in a row);
2. are injured;
3. display incorrect lifting technique;
4. put the Jerry cans down in the non-rest areas;
5. reach a distance of 600 metres.

Number of participants that can be tested simultaneously

4

Personnel

1. One PTI to take warm up and administer general but non-specific encouragement
2. 2 researchers to demonstrate task, administer test, measure, set up, record time and video tape. Recording environmental temperatures at start and finish

3. 1.82 m wall climb from running and standing starts

Task requirement

1. You are required to successfully climb over the 1.82 m wall from a running start then return over the wall from a standing start.
2. Whilst completing this task you are required to wear patrol order, ballistic vest and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg).

Purpose

1. The 1.82 m wall climb is a commonly undertaken trade task. It requires speed, leg power and upper body strength.
2. The purpose of the task is to simulate aspects of jumping, stair climbing and climbing in urban operational tasks.

Equipment requirements ADF	**Equipment requirements UB**
1. 1.82 m wall **2.** 2 witches hats	**1.** 2 stop watches (and splits) **2.** tape measure (20 metre and 3 metre) **3.** marking paint **4.** Recording sheets

Specifications

1. The test will require you to run ***carrying weapon (unslung)*** from the five metres witches hat to the wall jump/climb over the wall wait and reorganise yourself for five seconds and then return over the wall again. You will then be required to complete the task three times with a 10 minute break between each attempt for recovery.
2. You will have 30 seconds in which to achieve each wall climb.

Measures

1. The first measure is the time it takes to run and land over the wall with feet on ground. Then measure five seconds to reorganise yourself.
2. The second measure is the time it takes to get off the ground until you land over the other side.

Discontinuation criteria

You will be asked to discontinue the task when you:

1. cannot complete the climbing task in the required times;
2. are injured.

Number of participants that can be tested simultaneously

One

Personnel

1. One PTI to take warm-up and demonstrate task emphasising speed and technique
2. 2 researchers to administer test, measure, set up, record time and video tape. Recording environmental temperatures at start and finish

4. Leopard crawl

Task requirements

1. You are required to Leopard Crawl with correct technique for 25 m as fast as possible on a grassed surface
2. You will be required to wear patrol order, ballistic vest and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg).

Purpose

1. The purpose of the task is to simulate crawling through a 50 m tunnel.
2. Crawling in a prone position (known as Leopard Crawling) is a task commonly undertaken by infantry soldiers and ADG airmen.

Equipment requirements ADF	Equipment requirements UB
1. Oval (flat and mown) 2. 4 witches hats	1. 30 metre tape measure 2. Scales and boards × 2 3. Stop watches 4. Marking paint 5. Recording sheets

Specifications

You will be required to prop and drop from a standing start to begin the crawl. The finish is when your elbows are over the line (weapons are not to be slung). You will be required to complete the task twice.

Measures

The measure of the test will be the time taken to complete the 25 m crawl using a technique considered acceptable to the test administrator.

Discontinuation

You will be asked to discontinue the task if you:

1. become injured;
2. display incorrect crawling technique – one warning then stop.

Number of participants that can be tested simultaneously

Four

Personnel

1. One PTI to take warm up and demonstrate task emphasising speed and correct technique and assist ensuring quality control.
2. 2 researchers to administer test, measure, set up, record time and video tape. Recording environmental temperatures at start and finish

.

5. Urban rushing

Task requirements

1. You will be asked to perform constant velocity 22 m sprints and crouches.
2. You will be required to wear patrol order, ballistic vest and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg).

Purpose

1. The purpose of this task is simulate moving through a high threat urban environment requiring the infantry soldier/ADG to complete a series of sprint-recovery intervals that may be repeated many times.

Equipment requirements ADF	**Equipment requirements UB**
1. Oval (flat and mown)	1. Whistle 2. Stop watches 3. CD player with batteries 4. CD that works with player 5. 4 witches hats placed 22 metres apart on oval 6. Marking paint 7. Recording sheets 8. Tape measure (30 metres) 9. Scales and boards × 2

Specifications

1. The test will involve you performing externally paced 22 m sprints (at a constant velocity of seven seconds, as opposed to the increasing velocities of a shuttle run) interspersed with a standardised recovery period of seven seconds undertaken in a crouched position.
2. You will be required to cross the finishing line of each lap when the finishing beep or whistle sounds.
3. You will be required to return to the crouch position by the time the starting beep or whistle sounds.

Measures

The measure of the test is scored as the number of sprints completed

Discontinuation criteria

You will be asked to discontinue the test when you:

1. fail to complete two consecutive sprints at the required velocity.

Number of participants that can be tested simultaneously

Section

Personnel

1. One PTI to take warm up and assist ensuring quality control.
2. 2 researchers to administer test, measure, set up, weigh packs, record time and video tape. Recording environmental temperatures at start and finish

6. Infantry 10 km march and assault

Task requirement

1. In groups of two you will be required to complete a 10 km pre-fatiguing march on a road or track in marching order totalling 45 kg (webbing + pack + weapon) in 1 hr 50 min.
2. You are then required to complete a simulated assault course wearing patrol order, ballistic vest and helmet and carry a standard weapon (Steyr) (21.6 kg).

Purpose

The purpose of this task is to simulate:

1. A route march into an area of conflict
2. Conduct an assault on enemy

Equipment requirements ADF **10 km route march**	**Equipment requirements UB** **10 km route march**
1. Access to appropriate road and/or track.	**1.** Scales and boards × 2 **2.** 4 × two way radio **3.** Stop watches **4.** Recording sheets

Equipment requirements ADF **Assault**	**Equipment requirements UB** **Assault**
1. Oval or grassed area (flat and mown) **2.** 45 star pickets **3.** Fencing wire to create 70cm height entry hurdles and 100cm height hurdles on the exit **4.** Dolly or hammer to hit pickets in **5.** patrol order	**1.** Stop watches **2.** Marking paint **3.** 5 witches hats

Specifications

1. You will be required to walk on a set path in groups of two at five minute intervals. It is preferred that the path will enable us to provide you feedback. As you pass by the markers you will be provided with feedback regarding your pacing. If you are not within 15 seconds of the correct time you will be asked to stop until the correct time is reached. If you are too slow you will be provided feedback to catch up the time.
2. You will be required to march 5 km in 50 min, after which you will have a 10 minute break.
3. At the completion of the second 5 km march, a second 10 minute break will be taken .You will then take your ballistic vests and helmets out of your packs and put them on.
4. You are then required to complete the assault task in the fastest time possible. The order in which you do this will be the same order in which you completed the march.

Measures

The measure of the test is the time it takes to complete the assault course

Discontinuation criteria

You will be asked to discontinue the task when you:

1. are injured;
2. have reached a level of fatigue making you unable to complete the task;
3. fail to complete two check points at the required velocity (either too fast or too slow)

Number of participants that can be tested simultaneously

March - You will be sent off two at a time in five minute intervals
Assault – one participant at a time

Personnel

1. One PTI to demonstrate the assault task, lead warm-up and assist in ensuring quality control
2. 2 researchers to administer test, measure, set up, record time, weigh packs, provide feedback to soldiers, and video tape. Recording environmental temperatures at start and finish

7. Loaded incremental velocity run

Task requirement

1. You will be required to carry a constant load with increasing velocity on a circular oval in 20 m stages.
2. You are to wear basic uniform and carry a weighted pack and a standard weapon (Steyr) throughout the test (total weight carried is 22.5 or 12.5 or 10 kg for soldiers/airmen, IET and women respectively)

Purpose

The purpose of the test is to predict performance on loaded endurance tasks.

Equipment requirements ADF	Equipment requirements UB
1. Oval (flat and mown)	1. Witches hat × 10 2. Measuring wheel 3. CD player that can play CD (and batteries) 4. Recording sheets 5. Marking paint 6. Scales and boards × 2 7. Whistle 8. Tape measure (30 m)

Specifications

1. In groups of 4 you begin the test on the whistle and run to the next 20 m stage before the next whistle sounds.
2. Whistle will sound with decreasing time between sounds in a similar fashion as shuttle run, which you are familiar.
3. You will be allowed the first two stages of the run to get a feel for the tempo of the task before you are given any formal warnings.

Measures

The measure of the test is the level reached and the number of completed stages you can achieve within that level. In the case where you stop mid-stage your score will be the previous completed stage.

Discontinuation of criteria

You will be asked to discontinue the test when you:

1. miss 2 consecutive 20m timing gates (need to be within one metre of cone);
2. are injured.

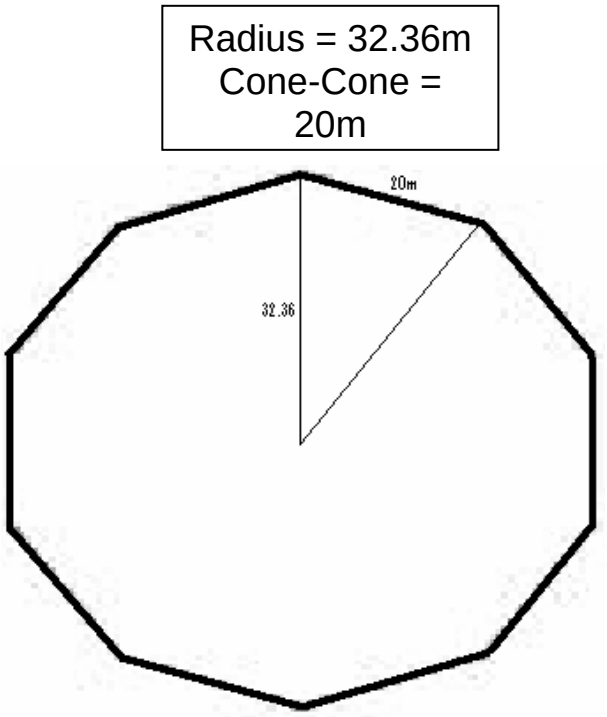
Number of participants that can be tested simultaneously

Between four and nine depending on staff.

Personnel

1. One PTI to take warm-up, demonstrate task and assist in quality control
2. 2 researchers to administer test, measure, set up, record time and video tape. Recording environmental temperatures at start and finish

Figure Pictorial representation of loaded incremental velocity run set-up



8. ADG pursuit and assault

Task requirement

1. You will be required to complete a 2.4 km pre-fatiguing pursuit.
2. You are then required to complete a simulated assault course.
3. You will be required to wear patrol order, ballistic vest and helmet and carry a standard weapon (Steyr) throughout the test (21.6 kg).

Purpose

The purpose of this task is to simulate:
1. A pursuit of enemy
2. patrol, assault and fight through

Equipment requirements ADF **2.4 km pursuit**	**Equipment requirements UB** **2.4 km pursuit**
1. Access to appropriate road and/or track	**1.** Stop watch **2.** Scales and boards × 2 **3.** Recording sheets

Equipment requirements ADF **Assault**	**Equipment requirements UB** **Assault**
1. Oval or grassed area (flat and mown) **2.** 45 star pickets **3.** Fencing wire or tape to create a 70cm on the entry hurdle and 100 cm on the exit hurdle **4.** Dolly or hammer to hit pickets	**1.** Stop watch **2.** Recording sheets

Specifications

1. You will be required to run on a set path at a pace of 9 km/hr, completing the pursuit in 16 minutes. The path that you take will enable us to provide you with feedback. As you pass by the markers we will write down your times and give you feedback on your pacing. If you are within 15 seconds of the correct time for that marker you will be allowed to continue on but if you are too fast you will be asked to mark time until the correct time is made.
2. At completion of the march a standard rest period of 10 minutes will be taken.
3. You are then required to complete the assault task in the fastest time possible. The order in which you do this will be in the same order in which you completed the march.

Measure

The measure of the test is the time it has taken to complete the assault course

Discontinuation criteria

You will be asked to discontinue the task when you:
1. are injured;
2. have reached a level of fatigue making you unable to complete the task;
3. fail to complete two check points at the required velocity (either too fast or too slow).

Number of participants that can be tested simultaneously

Pursuit- participants will be sent off one at a time with three minute intervals
Assault- one

Personnel

1. One PTI to take warm-up demonstrate assault and assist with quality control
2. 2 researchers to administer test, measure, set up, record time, provide feedback to soldiers, and video tape. Recording environmental temperatures at start and finish.

9. ADG 5km sustained patrol

Task requirement

1. In groups of two, you will be required to complete a 5 km march on a road or track in no more than one hour fifteen minutes.
2. The task requires you to wear marching order totalling 45 kg (webbing + pack + weapon.

Purpose

The purpose of this task is to simulate a patrol over an extended period into an area of conflict

Equipment requirements ADF	**Equipment requirements UB**
3. Access to appropriate road and/or track	**1.** Scales and boards × 2 **2.** 4 × two way radio **3.** Stop watches **4.** Recording sheets

Specifications

You will be required to walk on a set path in groups of two at five minute intervals. You will walk on a path that will enable us to provide you feedback on your pacing. As you pass by the markers we will write down your times and give you feedback on your pacing. If you are within 1.5 minutes of the correct time for that marker you will be allowed to continue on but if you are too fast you will be asked to mark time until the correct time is made.

Measures

The measure of the test is whether you successfully/fail to complete the ADG 5 km sustained patrol within in the required time.

Discontinuation criteria

You will be asked discontinue the task if you:

1. are injured;
2. have reached a level of fatigue making you unable to complete the task.

Number of participants that can be tested simultaneously

You will be sent of in pairs in five minute intervals

Personnel

1. One PTI to take warm-up
2. 2 researchers to administer tests, measure, set up, record time, weigh packs, provide feedback to soldiers, and video tape. Recording environmental temperatures at start and finish

Annex 10. Sample Recording Form

(After data transcription to computer spreadsheet; participants, test administrators and recorders de-identified)

Jerry Can Lift and Carry

Platoon:	XX
Date:	7/03/2005
Time:	11.00am
Recorder:	XX
Administrator:	XX

Platn	XX
Date	9/03/2005
Time	11.30 am
Rec	XX
Admn	XX

		Session 1			Session 2		
No.	**Name**	**Time**	**Dist**	**Notes**	**Time**	**Dist**	**Notes**
1			600			600	
2			600			600	
3			390			445	
4			275			295	
5			280			330	
6			600			550	
7			375			375	
8			600			482.5	Experiencing doms from JCC1, and Urban Rushing in morning
9							

Session 1 - 11.40 am	
Wet	19.2
Dry	23.1
BG	37.2
WBGTI	24.2
WBGTO	22.7
RH	37
HI	22
Flow	2

Session 2	
Wet	22.9
Dry	26.7
BG	40.6
WBGTI	28.6
WBGTO	27.2
RH	46
HI	27
Flow	0.8