DEVELOPMENT OF PHYSICAL EMPLOYMENT STANDARDS OF SPECIALIST PARAMEDIC ROLES IN THE NATIONAL AMBULANCE RESILIENCE UNIT (NARU)

Andrew G. Siddall[1], Mark P. Rayson[2], Ella F. Walker[1], Julianne Doherty[1], Josh I. Osofa[1], Tessa R. Flood[1], Beverley Hale[1], Steve D. Myers[1], Sam D. Blacker[1]

[1]Occupational Performance Research Group, University of Chichester, UK

[2]Mark Rayson Consulting Ltd, Bristol, UK

Corresponding author

Andrew G. Siddall

Research Fellow

Occupational Performance Research Group

University of Chichester

West Sussex, PO19 6PE

a.siddall@chi.ac.uk

ORCID: 0000-0002-3458-066X

This is an accepted article that has been peer-reviewed and approved for publication in Applied Ergonomics, but has yet to undergo copy-editing and proof correction. Please cite as an 'Accepted Article': DOI: 10.1016/j.apergo.2021.103460. The published version is available at the following link: https://doi.org/10.1016/j.apergo.2021.103460

Abstract

Aim: To develop evidence-based role-specific physical employment standards and tests for National Ambulance Resilience Unit (NARU) specialist paramedics.

Methods: Sixty-two (53 men, 9 women) paramedics performed an array of (1) realistic reconstructions of critical job-tasks (criterion job performance); (2) simplified, easily-replicable simulations of those reconstructions and; (3) fitness tests that are portable and/or practicable to administer with limited resources or specialist equipment. Pearson's correlations and ordinary least products regression were used to assess relationships between tasks and tests. Performance on reconstructions, subject-matter expert and participant ratings were combined to derive minimum acceptable job performance levels, which were used to determine cut-scores on appropriate correlated simulations and tests.

Results: The majority of performance times were highly correlated with their respective simulations (range of r: 0.73-0.90), with the exception of those replicating water rescue (r range: 0.28-0.47). Regression compatibility intervals provided three cut-scores for each job-task on an appropriate simulation and fitness test.

Conclusion: This study provides a varied and easily-implementable physical capability assessment for NARU personnel, empirically linked to job performance, with flexible options depending on organisational requirements.

Keywords: Physical employment standards, paramedics, fitness, occupational demands

1. Introduction

Many public service occupations, such as emergency responders and law enforcement, require workers to have a particular level of physical capability to perform their job safely and effectively (Gumieniak et al. 2013). Increasingly, physically demanding occupations are required to develop and adopt physical employment standards (PES) which are based on minimum acceptable performance of essential job-tasks (Tipton et al. 2012). Workers with inadequate physical ability to safely perform their role may increase risk to themselves, their colleagues and the public. Theoretically, by addressing mismatches between job demands and ability, appropriately-developed PES can seek to reduce injury risk, inform training, maintain operational effectiveness and promote safe and efficient work (Gebhardt 2019). Consequently, PES can form an important part of employers' *duty of care* and be implemented to select applicants or routinely test in-service personnel to evaluate their physical capability to perform their job roles. Research projects to develop evidence-based PES have been conducted in military personnel (Reilly et al. 2015; Sharp et al. 2017; Reilly et al. 2019), firefighters (Blacker et al. 2015; Siddall et al. 2016; Gumieniak et al. 2018) and law enforcement (Jamnik et al. 2010a) as well as other physically demanding jobs. However, there is a recognised sparsity of work concerning paramedics and disaster responders (Fischer et al. 2017; Lentz et al. 2019) who perform similar individual tasks to each of these more well-researched professions but require more targeted research.

Paramedics attend emergency incidents, such as road traffic/rail incidents, fires and critical illness to provide pre-hospital care to the public. Elements of these roles elicit high physical demand, particularly patient handling, stretcher manipulation and loading/unloading necessary equipment (Fischer et al. 2017). In the United Kingdom (UK), emergency situations that warrant more specialist capabilities are attended to by the National Ambulance Resilience Unit (NARU; Rue et al. 2019). The NARU works on behalf of each Ambulance Trust within the UK National Health Service and, by deploying specially trained paramedics, strengthens national resilience to a variety of challenging and hazardous emergency scenarios. As preparatory work for the present study, a task analysis of NARU

personnel identified a wide range of highly physically demanding essential tasks spread across three roles: Hazardous Area Response Team (HART); Chemical, Biological, Radiological and Nuclear (CBRN); and Marauding Terrorist Attack (MTA; Rue et al. 2019). These roles extend to treatment and/or extraction of casualties at height, in remote, urban and subterranean locations inaccessible by vehicle and from inland water, as well as performing mass decontamination and operating under a variety of terrorist threats.

Physical employment standards need to be judiciously designed and evaluated to ensure they are equitable, valid and unbiased (Tipton et al. 2012). Frameworks for PES development have been published to attempt to establish consensus on consistent methods and best-practice (Tipton et al. 2012; Jamnik et al. 2013; Reilly et al. 2015; Petersen et al. 2016). Broadly, these outline the stages of data capture to identify critical physically demanding tasks and standard operating procedures; define individual minimum acceptable performance; measure physical and physiological demand; select or design assessment tests and/or task simulations; and derive valid and appropriate standards. There are numerous challenges to measuring actual job performance given the complex nature of job-tasks performed by personnel attending emergency incidents and the potential subjectivity of "acceptable" performance. One established method is to use subject-matter experts to design realistic scenarios or "task reconstructions" which have high face validity (look like the criterion task) but (i) isolate the essential contribution that any one individual would have to make to tasks normally completed by a team and (ii) have controllable, measurable performance parameters to inform standard-setting (Blacklock et al. 2015). Measured performance on these task reconstructions by a group of personnel representative of the workforce can then be used to derive cut-scores using norm- or criterion-referencing, or a combination of both, of which various methods have been employed previously to try to balance objectivity and inherent subjectivity in standard-setting (Jamnik et al. 2013; Siddall et al. 2016). While the preferred approach to design job demands is to ergonomically fit demand to within the capabilities and limitations of a workforce, this is not always feasible for occupations that do not have static workplaces or predictable work scenarios. Even in complex scenarios, organisations are encouraged to review work practices and minimise the physical demands of job-tasks where possible,

and to make any ergonomic interventions that reduce demand available to all employees and review standards accordingly.

Complex scenario reconstructions support the development of evidence-based PES but are typically of prolonged duration and are resource-heavy, needing bespoke training locations, specialist equipment and support. In addition, these realistic scenarios require individual performers to be technically proficient (i.e. not pre-trained applicants). Using such reconstructions for employment testing is not always viable and, ultimately, PES must be implementable by an organisation to be effective (Reilly et al. 2015). Therefore, organisations often prefer simplified task simulations or gym-based physical assessments, particularly for testing applicants with lower technical proficiency. For implementation, it is valuable that these tests are easily replicable and of low financial cost while still being commensurate with the physical capability to perform the criterion tasks. It is typical, though, that as tasks become simplified for implementation, they proportionally lose job-relatedness and face validity. It is therefore valuable to understand the extent to which key components of fitness (e.g. muscular strength, muscular power, maximal aerobic power etc.) underpin the capability to perform work tasks in order to select valid and appropriate tests. Successful occupational task performance has been associated with a variety of gym-based fitness component tests such as time to cover set distances or incremental shuttle running as estimators of maximal aerobic power, vertical/broad jump tests for muscular power, and grip- and lift- strength tests (Hauschild et al. 2017). These associations support the notion that occupational capability tends to require a combination of upper- and lower- body strength and power, in addition to aerobic power and muscular endurance.

Currently, applicants to the HART role of NARU are required to successfully complete a physical competency assessment that was developed in 2007, and later refined in 2010, by a UK company (Optimal Performance Limited, Bristol, UK; reports not publicly available). Thereafter, the range of tasks and operating environments of HART personnel has widened. The job-task analysis by Rue et al. (2019) was conducted to inform current PES development for NARU, and to measure physiological demand during dedicated large-scale team scenarios (Rue et al. 2019). Therefore, the overarching objective of this investigation was to develop implementable evidence-based PES for the

three specialist NARU roles. Specifically, we aimed to examine the inter-correlation between performance of NARU personnel on (i) individual task reconstructions of the previously identified criterion tasks and (ii) both simplified simulations of those tasks and physical fitness tests, in order to support selection of appropriate assessments and associated cut-scores.

2. Methods

2.1 Task analysis

Preparatory work for this investigation was previously conducted and published, which comprised a job-task and physical demands analysis of NARU personnel and constituent roles (Rue et al. 2019). Briefly, this consisted of convening a subject-matter expert focus group to determine the critical physically demanding occupational tasks for NARU personnel and variation between different roles (HART, CBRN, MTA). This led to 11 identified criterion tasks, for which detailed realistic operational scenarios were developed (described in detail Rue et al. 2019) based on reasonable worst-case occupational requirements, of which the majority are typically performed as a team. These tasks were rescue tasks at height, from rubble, fast-flowing water and subterranean environments, and rescue over a distance and terrain inaccessible by vehicle. In addition, there were two vehicle unloading tasks; boarding a rigid inflatable boat (RIB) from water; and prolonged casualty treatment/decontamination/evacuation tasks in different forms of personal protective equipment (PPE): Gas-tight Suits (GTS), CBRN and MTA PPE.

2.2 Criterion task reconstruction and simulation development

In order to isolate and assess individual occupational performance to inform standard-setting in this investigation, we designed single-person "task reconstructions" for each previously identified criterion task. We conducted an ergonomic analysis of the criterion tasks and protocols, which included characterising the physical actions involved, distances travelled, equipment size/mass used and original team size. During protocol development, there was particular emphasis that the task reconstructions should: (i) adopt best, safe practice according to NARU standard operating procedures; (ii) reflect any

one individual's contribution to the original criterion task and; (iii) be reproducible in nature such that separate consecutive individuals would perform the task in the same manner and (iv) provide a simple performance measure (such as time to complete). To estimate individual contribution, as an example, for loading/unloading tasks the total number of items/equipment in the real-world scenario were categorised by how many personnel were required to carry them (and therefore their proportional mass), and by whether items were boxes or bags with handles, and also one- or two-handed carries. The number and types of these items were then replicated for the reconstructions, with the distances that one individual would cover in the team. In the event that any team task had a more physically demanding element performed by one individual (e.g., carrying a single-person item for the team), this task was included, as all personnel might be expected to perform the most physically demanding role of any team task.

Since task reconstructions are resource-heavy and difficult to reproduce for employment testing, we also explored simplified "task simulations" which aimed to predict task reconstruction performance by using the same physical characteristics as the tasks while being easily-replicable and practicable in different testing locations and/or using minimal/accessible equipment, resources and expertise. Our design criteria for simulations were that they could: (i) allow multiple individuals to be tested at once (ii) be conducted within a contained environment (such as a garage or hangar) or, if simulating a water-based task, be performed in a standard swimming pool; (iii) require minimal specialist equipment or clothing and; (iv) provide a simple performance measure (such as time to complete). The nature of some task reconstructions meant they were challenging to replicate and could not be adequately simulated under these criteria. For these tasks, we used predictive gym-based tests as surrogates for task simulations.

2.3 Study approach

The study was completed at two locations: Cardiff International White Water Centre (Cardiff, UK) for water-based tasks and tests, and the Defence CBRN Centre (Winterbourne Gunner, UK) for all other tasks and study elements. The project took place initially in September 2017 but due to testing

constraints, resources (time, personnel, reservation of training locations), and recruitment (accessibility to participants, work schedules), the sample size attained was insufficient and the study was replicated with different participants but identical methods and principal investigators in September 2019 to bolster the total sample size. It was important to consider that a systematic difference between two identical tasks on different collections may have signified inadvertent changes in protocols but also that it was likely, through random sampling, that averages and/or variation in personal characteristics and performances could contain some differences between cohorts. Data checks between the 2017 and 2019 cohorts showed similar representation in participant characteristics and high overlap between simulation and test performance scores, except for a few outlier performances (which were identified as caused by equipment being incorrectly set up, and are addressed at the beginning of the results section). Therefore, to maintain the focus of the study, both the 2017 and 2019 cohorts are presented as one combined cohort in the present study.

Participants attended the study testing for between 2-5 days (depending on their role) and were asked to complete:

- **Task Reconstructions:** Tasks with high ecological (real-world) validity performed in a realistic occupational setting with standard-issue equipment, which aim to replicate a single individual's contribution to a criterion task in order to assess occupational performance but would be challenging and resource-heavy to reproduce for employment testing.
- **Task Simulations:** Indoor tasks that employ minimal specialist equipment and space, which are simplified versions of one (or more) task reconstruction(s), designed to predict performance on task reconstructions and test physical fitness for occupational performance while being simple to perform if included in a PES testing battery.
- **Physical fitness tests:** A battery of fitness tests to measure components of fitness that may predict occupational task performance and would be simple to perform for NARU personnel if included in a PES testing battery.

Task reconstruction data were collected to represent actual occupational performance on critical job-tasks, from which minimum acceptable performance standards could be derived. Performance scores on predictive simulations and fitness tests could then be used to generate fitness cut-scores directly linked to job performance.

For all tasks and tests, a full verbal brief and demonstration was provided prior to commencement, and a project researcher and member of NARU directing staff moved with participants to provide supervision throughout. On any one trial day, no participants completed more than three task reconstructions. All task reconstructions, simulations and tests were separated by adequate recovery, the duration of which was related, primarily, to the duration of active work that each entailed (>2 hours between reconstructions; >45 min between simulations; >10 min between fitness tests). During the trial days, participants were allowed ad libitum access to food and drink.

2.4 Participants

Participants were approached by sending participant information documents through Ambulance Service Trusts. Sixty-two (53 men, 9 women) participants attended data collection, with role-specific representation of 28 for HART, 23 for CBRN and 24 for MTA (note: NARU personnel can fulfil more than one role simultaneously). Inclusion criteria were that participants were trained NARU personnel, currently operational and certified medically fit for duty. The latter was verified by completion of a medical screening questionnaire and, if applicable, a further medical evaluation from an on-site NARU medical advisor. Prospective participants were given a written brief in advance of the study and, on attendance at the data collection, received a verbal brief before giving written informed consent to take part. The study was approved by the University of Chichester Research Ethics Committee (Reference number: UOC REC 1718_54).

2.5 Preliminary measures

For all preliminary measures and fitness testing, participants were asked to wear lightweight gym clothes (shorts, t-shirt, trainers), but remove trainers for anthropometric measures. Participants were asked to refrain from smoking and eating in the hour preceding preliminary measures, and to not

consume alcohol or strenuously exercise in the preceding 12 hours. Height (Stadiometer, SECA Ltd, Birmingham, UK), body mass (weighing scales, SECA Ltd, Birmingham, UK) and estimates of body fat and fat-free mass via bio-electrical impedance (Tanita BC-418, Tanita Europe, The Netherlands) were collected on all participants. Body composition estimates from bio-electrical impedance were included solely to characterise the sample population, not for any predictive or clinical use.

2.6 Task reconstructions

We designed ten task reconstructions (Table 1) in order to realistically replicate the most physically demanding elements of the criterion tasks. These included seven tasks for HART personnel: Swift-water Rescue, Re-board RIB, Subterranean Rescue, Above-ground Rescue, Overground Rescue, Unload Incident Response Unit (IRU) and Movement in GTS; two tasks for the CBRN role: Unload Decon(tamination) vehicle and Clinical Decon (Recovery Role); and an MTA task for MTA-trained personnel. We omitted one task (the Over Rubble Rescue) described previously in the task analysis by Rue et al. (2019) after it was trialled because it could not adequately reflect the demands of the criterion task nor be adequately replicated. For each reconstruction, participants wore the appropriate Personal Protective Equipment (PPE) for that task (detailed in Table 1). Participants completed each task reconstruction individually and were encouraged to adopt best individual effort while maintaining safe operating procedures. Time to complete the task constituted the score on all reconstructions.

2.7 Task simulations

We designed eight task simulations to replicate the most physically demanding aspects of their respective task reconstructions, but in a simplified form (Table 1). The water-based simulations were completed in a standard indoor swimming pool which had one raised pool side 29 cm above water level. This was climate-controlled, with a water temperature of ~22°C. We designed the land-based simulations to be performed on a 10-12.5 m linear course, with each protocol stipulating a required number of shuttles during the task. Four parallel lanes (2.5-3.0 m wide) were set up in a large open-doored hangar, with marker cones at designated intervals to allow multiple participants to undertake simulations at the same time. Participants completed the simulations in standard uniform and, where

necessary, weighted vests and/or backpacks equivalent to the additional mass of the PPE worn for the appropriate task reconstruction (Table 1). The hangar was not climate-controlled, and the doors were kept open to maintain a similar temperature to the outdoor spaces, but with little to no solar load or wind flow. Participants completed each task simulation individually and were encouraged to adopt best individual effort while maintaining safe operating procedures. Time to complete each simulation constituted the performance score.

2.8 Fitness testing

The fitness tests employed were selected from a combination of our previous experience using field-expedient tests in occupational groups and a literature search for tests used in similar organisations. As indicators of upper and lower body power, respectively, participants performed a seated medicine ball throw and standing broad jump, with distance, recorded to the nearest cm, as the performance measure. For the medicine ball throw (described previously; Lockie et al. 2018), participants positioned themselves with their back against a wall with legs out straight ahead and were instructed, without their back losing contact with the wall, to perform a maximal two-handed throw of a 4 kg medicine ball forward from the centre of their chest. For the standing broad jump, participants were instructed to perform a two-footed jump forward as far as possible, landing on two feet.

To indicate static maximal strength of the upper and lower body, respectively, we measured peak isometric hand-grip and upright pull strength using portable dynamometers (Takei, Japan), with kg (of force) measured to 0.1 of a kg, as the outcome measure. For hand-grip, after adjusting the dynamometer to individual hand size, participants were asked to grip the handle as forcefully as possible for three-five s with their arm extended by their side. For the upright pull (described previously; Coldwells et al. 1994), the dynamometer has a height-adjustable handle affixed to a baseplate on which participants stood with a hip-width stance. Participants were asked to bend marginally at the hips and knees when the handle was adjusted to approximately mid-thigh position. Participants were then instructed, while maintaining a straight back, to pull forcefully upwards on the handle for three-five s. For all strength tests, participants were given three attempts.

For agility and speed, respectively, the agility T-drill and a 60-m sprint were performed, with time to complete the tasks, recorded to the nearest 0.1 of a s, used as the performance scores. For the T-drill, participants are required to run forward 10 m to a central cone, then move laterally (side-stepping) 5 m to a left-hand cone, move laterally 10 m in the opposite direction to a right-hand cone, move laterally back to the central cone and then run backwards to the start point as fast as possible. In this test, participants are instructed to touch each cone with their hand for the test to be valid. The 60-m sprint was performed outdoors on tarmac, and split times were taken at 10 and 20 m, as well as for the total distance.

For an estimate of maximal oxygen uptake ($\dot{V}O_2$ max), a multi-stage fitness test was performed using a 20-m shuttle course on tarmac and an audio signal protocol from a dedicated phone application (Bleep Test Lite, Bitworks Design, UK) played via a speaker. Participants continued the test until volitional fatigue, or from not maintaining pace with the audio signals for more than two consecutive shuttles. The multi-stage fitness test score was recorded as the total number of shuttles completed by each participant and was used to estimate $\dot{V}O_2$ max (relative to body mass; units: $mL \cdot kg^{-1} \cdot min^{-1}$) using the equation derived previously (Ramsbottom et al. 1988).

For personnel in the HART role, a 25 m swim test was also completed, with time to complete to the nearest 0.1 of a s used as the performance score. For the swim test, participants wore regular swim wear and were asked to keep their head out of the water during the test, as if keeping eyesight on a casualty.

Table 1. Task reconstruction and simulation descriptions for each NARU role

	Descriptions	
Role	**Task Reconstruction**	**Task Simulation**
HART	**Swift-water rescue** – participants (in dry suit and swift-water rescue PPE; ~9 kg), swam tethered across fast-moving water* to a marker, simulating a casualty, 25 m from the start point.	**Resisted swim** – participants (in normal swim attire), swam 25 m in an indoor swimming pool while wearing a resistance parachute (29 cm, FINIS Ltd, USA).
	Re-board RIB – participants (in dry suit and swift-water rescue PPE; ~9 kg) climbed aboard a RIB from the water, unaided by colleagues, but using a webbing foot-strop (an adjustable length strap with a loop used as a foot-hold) for assistance.	**Low & high pool exits** – participants (in normal swim attire), starting ~1 m from the poolside, exited the pool unaided, to standing, onto the pool edge at water level (**low**) and a raised pool edge ~29 cm above water level (**high**).
	Subterranean Rescue – participants (in Urban Search & Rescue PPE; ~8 kg) moved 60 m through a confined crawl-way, over and under obstacles, while carrying a "Tombstone" kit bag (~14 kg). On reaching a simulated casualty strapped to a multi-integrated body-splint (total mass 44 kg; representing half of a casualty), participants then removed the Tombstone and low-dragged (keeping head below window-line) the casualty 60 m over and under obstacles through a train carriage. This was completed two further times for a total of three circuits.	**SIM$_{SUB}$** – participants (in Urban Search & Rescue PPE; ~8 kg) crawled 60 m (3 x shuttle around 10 m marker cone), affixed straps on multi-integrated body-splint around a simulated casualty and low-dragged the casualty (total mass 44 kg) 60 m (3 x shuttle around 10 m marker cone). After a fixed recovery of 3 min, participants low-dragged a second, heavier casualty on a multi-integrated body-splint (total mass 88 kg) 5 m to replicate the end of the Above-ground rescue task.
	Above-ground rescue – participants (in "Safe Working at Height" PPE; ~14 kg and carrying a "Tombstone" kit bag ~14 kg), ascended a 7 m vertical ladder to a platform, placed the Tombstone on the platform in order to move between scaffolding bars onto the platform, donned the Tombstone again and descended to ground level via three shorter ladders. This was completed a total of 10 times to simulate an ascent/descent of a 70 m structure. Participants then low-dragged a casualty on a multi-integrated body-splint (total mass 88 kg) 5 m to complete the task.	*No specific simulation: The physical demand of vertical climbing could not be replicated within simulation criteria. However, the final casualty drag was included at the end of HART SIM$_{SUB}$(terranean).*
	Overground rescue – participants (in incident ground kit; ~8 kg) traversed a pre-set 3 km undulating course on foot. At the end of the route, participants lifted a weighted bag (22 kg; representing a quarter of a casualty) onto a table to simulate the rear of an extraction vehicle.	*No specific simulation: The terrain and prolonged walk/job could not be adequately replicated under simulation criteria. However, performance on the multi-stage fitness test was considered a field-expedient proxy for the capacity to perform a 3 km run, and lifting of casualties was included in the HART SIM$_{GTS}$.*
	Unload IRU – participants (in incident ground kit; ~8 kg) lifted and carried a representative proportion (one quarter) of the total equipment carried in an IRU vehicle 15 m, then a smaller number of items 200 m to replicate a casualty clearing area. Participants (in any order) carried 18 x 15 kg bags, 17 x 15 kg boxes, 5 x 20 kg boxes and 1 x 25 kg box 15 m (to a 7.5 m marker and back), placing the item(s) on the ground between each shuttle. Each shuttle was separated by an unladen 15 m walk. Six items (2 x 15 kg boxes, 4 x 15 kg bags) were carried 200 m (to a 100 m marker and back), separated by an unladen	**SIM$_{IRU}$** – participants (in incident ground kit; 8 kg) alternated between walking with and without a weighted box. The same box was used and weights were added/removed when applicable. Participants carried (in this order) 7 x 15 kg boxes, 2 x 20 kg boxes and 1 x 25 kg box each 15 m (to a 7.5 m marker and back) placing the box down on the ground between each shuttle. Each shuttle was separated by an unladen 15 m walk. The 15 kg box was then

	200 m walk. For this task, participants were permitted to carry two bags at once to reduce total shuttles but boxes had to be carried individually.		carried 200 m (to a 10 m marker and back 10 times) twice, separated by an unladen 200 m walk.
	Movement in GTS – participants (wearing extended duration breathing apparatus and GTS; ~37 kg) pushed an unladen wheeled litter 200 m (to a 100 m marker and back), followed by alternating between 2 x 20 m walks (to a 10 m marker and back), and 2 x 20 m forward-facing drags of a casualty on a "Team 8" stretcher (total mass 41 kg). Participants then lifted a weighted bag (41 kg; simulating lifting half the casualty) from the ground onto the litter. Participants then completed exactly 5 min of chest compressions on a training mannequin before wheeling the (now) laden litter 200 m (to 100 m marker and back) to end the task.		**SIM$_{GTS}$ –** participants (wearing incident ground kit (8 kg), a weighted vest (7 kg) and a weighted day sack (22 kg); total ensemble: ~37 kg) walked 200 m (to a 10 m marker and back 10 times) before alternating between 2 x 20 m walks and 2 x 20 m forward-facing drags of a casualty on a "Team 8" stretcher (total mass 41 kg). Following the drags, participants lifted a 41 kg weighted bag onto a table, then lowered the bag to the ground under control. Participants finished the task with a final 200 m walk.
CBRN	**Unload decon vehicle** – participants (in incident ground kit; ~4 kg) lifted and carried a representative proportion (one sixth) of the total equipment load of a decon vehicle 25 m. Participants carried 5 x 15 kg boxes, 1 x 20 kg box, 2 x 25 kg bags, 2 x 25 kg boxes 25 m (to a 12.5 m marker and back), placing the item(s) on the ground between each shuttle. Each shuttle was separated by an unladen 25 m walk/jog. Each item had to be carried individually.		**SIM$_{UDECON}$** – participants (in CBRN clothing; ~4 kg) alternated between walking with and without a weighted box to a 12.5 m cone and back. The same box was used and weights were added when applicable. Participants carried 5 x 15 kg boxes, 1 x 20 kg box and 4 x 25 kg boxes 25 m. Each shuttle was separated by an unladen 25 m shuttle.
	Clinical decon (recovery role) – participants (wearing Personal Respiratory Protective Suit; ~12 kg) pushed an unladen litter 200 m (to 100 m marker and back). Participants then lifted a weighted bag (41 kg; simulating lifting half the casualty) onto the litter and secured it with a central strap before pushing the (now) laden litter 200 m (to 100 m marker and back). On return, participants unhooked the strap on the litter and lowered the weighted bag to the ground under control. This circuit was performed three times to simulate recovering three casualties.		**SIM$_{CDECON}$** – participants (in CBRN clothing (~4 kg) and a weighted vest (8 kg); total ensemble ~12 kg) pushed an unladen wheeled litter 200 m (to a 10 m marker and back 10 times) before lifting a 41 kg weighted bag onto a table. Participants then pushed a laden (41 kg) litter 200 m before lowering the weighted bag from the table under control. This circuit was completed three times.
MTA	**MTA task** – participants (in full ballistic PPE; ~19 kg) completed five circuits of approximately 900 m, with each circuit comprising multiple activities in the same order. In each circuit: Participants first completed a 400 m approach walk/run, then repeated a casualty triage sequence three times (20 m dash to-, triage- (replace tourniquet) and drag-casualty (mass 83 kg) 10 m). Participants then triaged and dragged four more 83 kg casualties 5-10 m. Participants then sprinted 60 m to an area of cover. A 44 kg casualty on a "Team 8" skid was then dragged 200 m (to a 100 m marker and back) before participants walked back to the start point.		**SIM$_{MTA}$ –** participants (in MTA clothing (5 kg) and weighted vest (14 kg); total ensemble: ~19 kg) performed a 400 m walk (to a 10 m marker and back 20 times) followed by a casualty triage sequence seven times, three with sprints and four with walks (sequence: sprint/walk around 10 m marker and back, place and undo a tourniquet on a casualty, drag casualty around a 5 m marker and back). Participants finished the tasks with a 60 m sprint (to a 10 m marker and back three times) and a final 400 m walk.

Note: HART=Hazardous Area Response Team, CBRN=Chemical, Biological, Radiological, Nuclear, MTA=Marauding Terrorist Attack, PPE=Personal protective equipment, RIB=Rigid inflatable boat, IRU=Incident Response Unit, GTS=Gas-tight Suits, SIM=simulation. *6 $m^3 \cdot s^{-1}$.

2.9 Measurement of physiological strain and physical demand

We measured participants' physiological strain at 5-s intervals by chest-mounted heart rate (HR) monitor (Polar Team 2, Polar Electro Ltd, Finland) during all task reconstructions, task simulations (except water-based elements) and the multi-stage fitness test. For each participant, the highest HR observed during the entire testing period or the multi-stage fitness test (whichever was higher) was used to express activity intensity as a percentage of HR max (%HRmax). We then calculated the percentage of each task spent within or above "Hard" intensity activity (Hard-to-Very Hard threshold: 77% HR max) using intensity zones described elsewhere (Howley 2001). We also asked participants to rate the overall task physical demand on a 1-6 scale ("Very Light" – "Maximum", not yet validated) at the end of each reconstruction and simulation. This scale was used rather than a typical rating-of-perceived exertion scale as the question was related to a job-task as a single entity rather than current real-time exertion.

2.10 Task minimum acceptable performance standard opinion

To inform standard-setting, after each task reconstruction and simulation, we informed participants of their performance time and asked, "Bearing in mind your level of fitness, where would you consider the minimum acceptable pass-standard to be?". After each task simulation, we also asked participants how representative the physical demand of the simulation was of the equivalent task reconstruction on a 6-point scale ranging from 1, "not at all," to 6, "very well". Additionally, during the investigation, experienced NARU directing staff were observing all tasks and tests. These personnel were all instructors at the Defence CBRN Centre each with more than 15 years of experience in the organisation, and therefore were (or had historically been) responsible for training and assessment of these technical skills. These staff were not direct supervisors of any participants involved in the study. At the conclusion of each battery of task reconstructions and simulations, we asked these subject-matter expert staff individually for their opinion on what they would consider the minimum acceptable performance time for each exercise. At least two instructors were asked for their opinion on each task, and each staff member had observed all participants for the tasks they rated.

2.11 Data analysis

All statistical analyses were performed using either researcher-produced spreadsheets (Microsoft Excel, Microsoft Ltd, USA) or Statistical Package for the Social Sciences Version 23 (SPSS; IBM, USA). Unless otherwise stated, summary data for participant characteristics, task and simulation performance and physiological demand are mean ± standard deviation (SD). For categorical or ordinal data such as questionnaire scales or non-continuous data, median or mode with range have been adopted. On these subjective scales, the mode has been adopted where we deemed that the majority "vote" was more informative for characterising an overall task, and not analysed statistically. Pearson's correlation coefficients (r) were used to assess the strength of relationship between performances on task reconstructions and a) individual task simulations and, b) anthropometry and fitness tests. As it is not possible to achieve full control of error in human field trials, *P*-values for associations were computed to assess compatibility between the observed data and the entire underlying test model (i.e. the combination of the null hypothesis of no association and the collective statistical assumptions used to compute the *P*-value). This is such that a lower observed p would indicate lower compatibility of the data with the null test model. No specific long-run error rate thresholds were set, to allow readers to interpret data based on their own error costs.

For the purpose of standard-setting, we employed ordinary least products (OLP) regression to produce a regression line between task reconstruction score and performance on the simulations and fitness tests that were found to be suitably strongly correlated. The OLP method is similar to the more commonly used ordinary least squares analysis but rather than allowing random error on only one axis, OLP allows random (i.e. measurement) error in both x and y variables (Ludbrook 2012; Wilkinson et al. 2014). Importantly, we adopted this technique because in the present study random error could occur in task or simulation measurement. Additionally, by combining the error observed on both axes, this method allows for prediction in either direction. This approach was important because the derived relationships may later be used by NARU to (i) predict simulation performance time from a given task standard or (ii) determine a task duration from a given simulation time. For each OLP regression line, we computed residual standard error (RSE) as a measure of the variability in the prediction as well as

70%, 80% and 90% compatibility ("confidence") intervals (CIs) on the y-axis specifically for standard-setting procedures.

2.12 Standard-setting procedures

Developing PES that are evidence-based, and specifically based on the minimum acceptable performance of essential job-tasks, underpins standards being fair and defensible. As such, it was imperative that our analysis methods made an empirical link between acceptable job performance on criterion task reconstructions and any resultant predictor tests or simulations (and their associated cut-scores). The process for standard-setting for physically demanding occupations has been reviewed extensively (Tipton et al. 2012; Jamnik et al. 2013) and inevitably contains a level of subjectivity (e.g. what is minimally acceptable/successful performance) within objective data. For each task reconstruction, we derived a proposed "minimum acceptable performance" standard by triangulating three sources of evidence: (1) The mean plus 1 SD of the performance time of the participants that completed the task reconstruction (adopted previously; Jamnik et al. 2010), with the rationale that ~83% of the workforce sample is incorporated into the calculation of a physical employment standard; (2) the minimum acceptable performance standards proposed by the NARU subject-matter experts that observed the reconstructions and; (3) the minimum acceptable performance standards proposed by participants after completing each reconstruction. The latter two sources were incorporated in order to not rely solely on norm-referencing from best-effort performances, to include expert opinion and to attempt to balance objectivity and subjectivity. We calculated an arithmetic mean of these three sources of evidence to derive a minimum acceptable standard for each task reconstruction. This analysis was completed on all participants who completed the task reconstructions and simulations irrespective of whether performance times would be deemed unacceptable by one or more methods retrospectively.

For the purpose of presenting several possible cut-scores to the organisation in question, for a given task reconstruction standard, we used the upper bound of each compatibility interval (70%, 80%, 90%) of the OLP regression line to predict three corresponding scores on the simulation or fitness test.

In this approach, wider CIs mean the possible range of predictions is more compatible with the observed data, but this results in a cut-score that is harder to achieve (on the simulation) and setting a lower level of compatibility results in a cut-score that is easier to achieve. The option of multiple cut-scores was to allow the organisation discretion on different levels of leniency for standard-setting (e.g. for borderline categories or to control inflow of applicants), while still maintaining an empirical link to criterion performance.

3 Results

3.1 Participants

All participants completed each physical fitness and anthropometric test except for one HART participant who was unavailable for the swim tests. All participants successfully completed the task reconstructions and simulations that they attempted. Performances for three participants on the Re-board RIB task reconstruction were removed as outliers because the foot-strop (a strap with a loop for a foot-hold which is attached to the side of the RIB) was set to the incorrect length, which substantially delayed completion of the task, and was only rectified after their performances. Five participants in the HART cohort sustained minor injuries during testing that prevented them from performing certain land-based HART simulations. Specifically, these were classified on site as three truncal musculoskeletal overuse injuries and two acute elbow injuries sustained from impacts with objects. Participant characteristics and physical fitness test performance data are organised in Table 2.

Table 2. Participant descriptive characteristics and physical fitness test data organised by NARU role. Data are mean (±SD).

Variable	Role			All
	HART	CBRN	MTA	
n (Male/Female)	28 (23/5)	23 (19/4)	24 (21/3)	62 (53/9)
Age (y)	39 (±8)	41 (±11)	38 (±10)	40 (±10)
Height (m)	1.77 (±0.08)	1.77 (±0.10)	1.76 (±0.08)	1.77 (±0.08)
Body mass (kg)	83.8 (±14.9)	87.4 (±17.0)	84.0 (±16.1)	85.0 (±15.1)
Estimated body fat (%)	21.4 (±7.3)	24.6 (±6.8)	21.6 (±7.2)	22.2 (±7.4)
Estimated FFM (kg)	65.5 (±11.1)	65.6 (±11.9)	65.3 (±10.1)	65.8 (±10.6)
BMI ($kg \cdot m^{-2}$)	26.7 (±3.6)	27.8 (±3.5)	27.0 (±4.1)	27.1 (±3.7)
Estimated $\dot{V}O_2$ max ($L \cdot min^{-1}$)	3.48 (±0.75)	3.14 (±0.83)	3.21 (±0.70)	3.31 (±0.77)
Estimated $\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	41.7 (±7.3)	36.2 (±8.1)	38.7 (±7.1)	39.3 (±8.0)
Handgrip (kg)	51 (±11)	50 (±11)	50 (±10)	50 (±11)
Upright pull (kg)	182 (±41)	195 (±47)	193 (±47)	185 (±42)
Standing broad jump (m)	1.96 (±0.30)	1.84 (±0.31)	1.95 (±0.28)	1.93 (±0.30)
Medicine ball throw (m)	4.50 (±0.74)	4.55 (±0.77)	4.52 (±0.60)	4.53 (±0.69)
Agility T drill (s)	13.8 (±1.6)	14.2 (±1.9)	13.4 (±1.5)	13.8 (±1.7)
60 m Sprint (s)	9.9 (±1.2)	10.5 (±1.4)	10.0 (±1.0)	10.1 (±1.2)
Swim 25 m (s)	21.2 (±4.3)	N/A	N/A	21.2 (±4.3)

Note: HART=Hazardous Area Response Team; CBRN=Chemical, Biological, Radiological, Nuclear, MTA=Marauding Terrorist Attack; FFM=Fat free mass; BMI=Body mass index. Estimated body fat and FFM via bioelectrical impedance.

3.2 Task reconstructions and simulations

The task reconstructions ranged in duration (mean ± SD) from 23 (± 6) s (swift-water rescue) to 1 hour 24 (± 13) min (MTA task) (Table 3). With the measures of physical demand taken together, the most arduous task reconstructions were Overground rescue and Unload IRU for HART personnel, and Clinical decon (recovery role) for CBRN personnel. The simulations ranged from 5 (± 1) s (Low pool exit) in duration to 17 min 18 s (± 2 min 29 s) (SIM_{CDECON}).

Table 3. Completion time and physical demand parameters of task reconstructions (titled bold) and their respective simulations. Data are mean (±SD) or mode and range, as specified.

Role	**Reconstruction** - Simulation	Time to complete (mm:ss)	Mean HR (%HRmax)	Time in ≥ hard zone (% of task)	Perceived physical demand (Mode (range))
HART	**Swift-water rescue**	00:23 (±00:06)	-	-	3 (2 - 4)
	- Resisted swim	00:25 (±00:05)	-	-	3 (2 - 4)
	Re-board RIB	00:22 (±00:14)	-	-	3 (1 - 4)
	- Low pool exit	00:05 (±00:01)	-	-	1 (1 - 3)
	- High pool exit	00:06 (±00:02)	-	-	3 (1 – 5)
	Subterranean rescue	30:03 (±11:05)	85 (±11)	78 (±38)	5 (5 - 6)
	- SIM_{SUB}	07:53 (±01:20)	73 (±8)	42 (±31)	4 (2 - 5)
	Above-ground rescue	34:17 (±09:00)	81 (±8)	74 (±29)	5 (3 - 6)
	Overground rescue	22:10 (±02:45)	87 (±4)	91 (±13)	4 (3 - 5)
	Unload IRU	28:20 (±05:45)	86 (±8)	92 (±12)	4 (3 - 6)
	- SIM_{IRU}	11:25 (±02:31)	77 (±7)	63 (±33)	3 (3 - 5)
	Movement in GTS	12:34 (±01:33)	80 (±5)	70 (±24)	4 (3 - 6)
	- SIM_{GTS}	05:52 (±00:45)	78 (±8)	66 (±23)	3 (2 - 4)
CBRN	**Unload decon**	05:07 (±00:47)	83 (±8)	83 (±20)	3 (3 - 5)
	- SIM_{UDECON}	04:50 (±00:47)	81 (±9)	72 (±35)	3 (3 - 5)
	Clinical decon	16:38 (±01:47)	87 (±6)	90 (±21)	5 (3 - 6)
	- SIM_{CDECON}	17:18 (±02:29)	79 (±11)	66 (±36)	4 (2 - 6)
MTA	**MTA**	84:09 (±13:02)	85 (±5)	86 (±20)	5 (4 - 6)
	- SIM_{MTA}	16:29 (±01:43)	77 (±9)	64 (±30)	4 (3 – 5)

HART=Hazardous Area Response Team; CBRN=Chemical, Biological, Radiological, Nuclear, MTA=Marauding Terrorist Attack; RIB=Rigid Inflatable Boat; IRU=Incident Response Unit; GTS=Gas-tight Suits; SIM=simulation.

All land-based simulation performance times were highly correlated (coefficients ranging from 0.73-0.90) with their respective task reconstruction times ($p<0.05$; Table 4). Associations between water-based simulations and tasks were less strong, indicated by a moderate correlation between resisted swim and swift-water rescue ($r=0.47$; $p=0.02$) and low-moderate correlations between low ($r=0.38$, $p=0.08$) and high ($r=0.28$; $p=0.15$) pool exits with Re-boarding the RIB. Participant ratings (mode (range)) on the extent of similarity between demands of reconstructions and simulations ranged from 'moderate' (3 (1-5)) for low pool exit to 'well' (5 (3-6)) for SIM_{IRU}, where all land-based simulations demonstrated a majority of "moderately well" or higher. In the HART task reconstructions that did not have dedicated simulations, Above-ground rescue had the strongest association with

medicine ball throw (r=-0.64, p<0.001) and Overground rescue with (estimated) relative $\dot{V}O_2$ max (r=-0.77; p<0.001). The majority of task reconstructions were correlated, to a moderate-high level, with five or more physical fitness parameters such that higher physical fitness equated to lower (quicker) performance time on task reconstructions (Table 4). Swift-water rescue was the only task reconstruction to not have at least one correlation with a fitness parameter with r>0.6 (strongest correlation: 25-m swim, r=0.39, p=0.049).

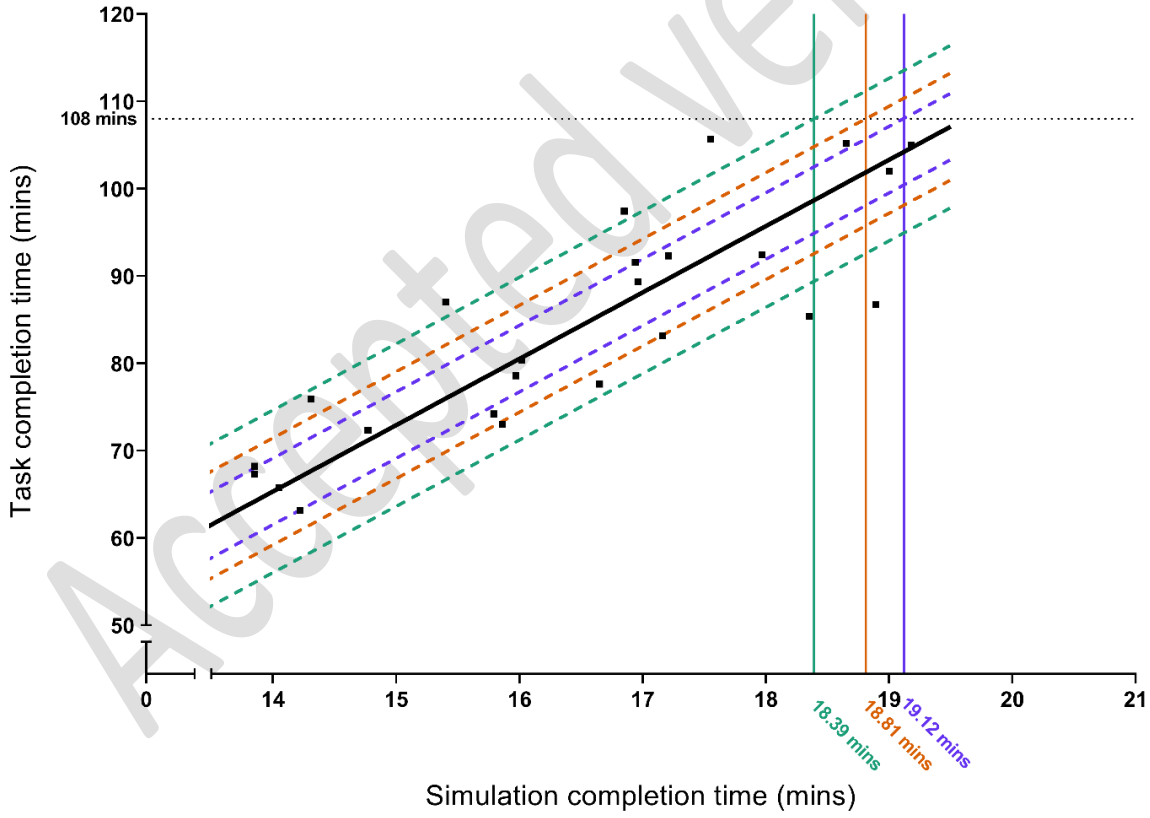
Table 4. Correlation coefficients (*r*) for relationships between performance on each task reconstruction, respective simulations, and physical fitness tests.

	Role									
	HART							**CBRN**		**MTA**
Fitness parameter	Swift-water rescue	Re-board RIB	Subterranean rescue	Above-ground rescue	Overground rescue	Unload IRU	Movement in GTS	Unload decon	Clinical decon	MTA
Estimated $\dot{V}O_2$ max ($L \cdot min^{-1}$)	-0.21	-0.78**	-0.58**	-0.56**	-0.41*	-0.69**	-0.70**	-0.70**	-0.72**	-0.68**
Estimated $\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	-0.15	-0.68**	-0.68**	-0.56**	-0.77**	-0.50**	-0.43*	-0.79**	-0.77**	-0.73**
Hand grip (best)	-0.16	-0.71**	-0.56**	-0.51**	-0.32	-0.53**	-0.55**	-0.45*	-0.38	-0.20
Upright pull	-0.23	-0.77**	-0.54**	-0.51**	-0.32	-0.52**	-0.45*	-0.31	-0.48*	-0.14
Standing broad jump	-0.29	-0.63**	-0.57**	-0.54**	-0.43*	-0.55**	-0.60**	-0.70**	-0.62**	-0.43*
Medicine ball throw	-0.35	-0.62**	-0.61**	-0.64**	-0.28	-0.67**	-0.76**	-0.61**	-0.50*	-0.22
Agility T drill	0.17	0.39	0.46*	0.47*	0.45*	0.54**	0.59**	0.82**	0.74**	0.48*
60-m sprint	0.19	0.70**	0.51**	0.54**	0.47*	0.67**	0.60**	0.73**	0.76**	0.40
25-m swim	0.39*	0.49*	0.59**	0.35	0.38	0.39	0.27	-	-	-
Simulation	0.47*	0.38 (Low) 0.28 (High)	0.75**	na	na	0.79*	0.76**	0.90**	0.73**	0.85**

*Note: HART=Hazardous Area Response Team; CBRN=Chemical, Biological, Radiological, Nuclear, MTA=Marauding Terrorist Attack; RIB=Rigid Inflatable Boat; IRU=Incident Response Unit; GTS=Gas-tight Suits. *p<0.05 **p<0.01.*

3.3 Standard-setting

The initial task reconstruction performance standards, triangulated from participant performances (mean + SD), and subject-matter expert- and participant- opinion, are shown in Table 5 (column 2). From each of these performance standards, we produced three levels (A, B and C) of cut-scores on simulations and fitness parameters from the intersection with the 70%, 80% and 90% CIs of each OLP regression line (example Figure 1; applying performance standard derived for the MTA task: 108 min [1 hr 48 min]). Table 5 also presents OLP regression parameters with resultant cut-scores on associated simulations and tests. Data and cut-scores are presented for each simulation that achieved a moderate-strong correlation with its corresponding task reconstruction and low compatibility with the null model, and for the fitness test(s) that were most highly correlated with each reconstruction.



[COLOUR PRINTING REQUIRED] **Figure 1.** Example plot of individual performance times on a task (MTA, the derived minimum acceptable performance standard: 108 min [1 hr 48 min]) and simulation (SIM_{MTA}) with OLP regression line (black solid line) and 90% (green dashed), 80% (orange dashed) and 70% (purple dashed) compatibility (confidence) intervals. Horizontal grey dotted line denotes a notional pass standard on a task intersecting with each upper compatibility level to derive a predicted performance cut-score for the corresponding simulation (90%: green solid line; 80%: orange solid line, 70%: purple solid line). Note that for resultant simulation cut-scores, the graphical output displays decimal minutes, not min:ss.

Table 5. Minimum acceptable task reconstruction standards, regression information and three levels (A: derived from 90% CI; B: 80% CI, C: 70% CI) of regression-derived cut-scores on associated simulations and fitness tests.

Task reconstruction	Task reconstruction standard	Simulation/ Fitness test	OLP regression				Simulation/test cut-scores
			Slope	Intercept	RSE		
Swift-water rescue	29 s	Resisted swim	1.127	-0.093	0.098	A	24 s
						B	26 s
						C	28 s
		25-m swim	0.021	-0.077	0.105	A	20 s
						B	22 s
						C	24 s
Re-board RIB	31 s	$\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	-0.018	1.060	0.102	A	38.0 $mL \cdot kg^{-1} \cdot min^{-1}$
						B	35.5 $mL \cdot kg^{-1} \cdot min^{-1}$
						C	33.7 $mL \cdot kg^{-1} \cdot min^{-1}$
		Upright pull	-0.003	0.866	0.086	A	153 kg
						B	140 kg
						C	131 kg
Subterranean rescue	47 min 29 s	SIM_{SUB}	8.845	-39.495	8.465	A	8 min 36 s
						B	9 min 2 s
						C	9 min 20 s
		$\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	-1.527	93.708	9.013	A	37.8 $mL \cdot kg^{-1} \cdot min^{-1}$
						B	35.3 $mL \cdot kg^{-1} \cdot min^{-1}$
						C	33.4 $mL \cdot kg^{-1} \cdot min^{-1}$
Above-ground rescue	46 min 5 s	Medicine ball throw	-12.686	90.771	7.834	A	4.3 m
						B	4.0 m
						C	3.8 m
		$\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	-1.195	84.084	8.632	A	41.1 $mL \cdot kg^{-1} \cdot min^{-1}$
						B	37.9 $mL \cdot kg^{-1} \cdot min^{-1}$
						C	35.6 $mL \cdot kg^{-1} \cdot min^{-1}$
Overground rescue	28 min 58 s	$\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	-0.379	37.989	1.895	A	30.2 $mL \cdot kg^{-1} \cdot min^{-1}$
						B	28.0 $mL \cdot kg^{-1} \cdot min^{-1}$
						C	26.4 $mL \cdot kg^{-1} \cdot min^{-1}$
Unload IRU	41 min 2 s	SIM_{IRU}	2.456	0.337	4.054	A	14 min 27 s
						B	15 min 11 s
						C	15 min 42 s
		$\dot{V}O_2$ max ($L \cdot min^{-1}$)	-7.849	55.325	4.612	A	2.6 $L \cdot min^{-1}$
						B	2.3 $L \cdot min^{-1}$
						C	2.1 $L \cdot min^{-1}$
Movement in GTS	15 min 44 s	SIM_{GTS}	2.177	-0.159	1.175	A	6 min 37 s
						B	6 min 51 s
						C	7 min 1 s
		Medicine ball throw	-2.071	21.854	1.102	A	3.6 m
						B	3.4 m
						C	3.2 m
Unload decon	6 min 30 s	SIM_{UDECON}	1.002	0.280	0.358	A	5 min 45 s
						B	5 min 54 s
						C	6 min 1 s
		Agility T-drill	0.412	-0.750	0.484	A	16.1 s
						B	16.6 s
						C	17.0 s
Clinical decon	18 min 50 s	SIM_{CDECON}	0.909	1.213	1.336	A	17 min 30 s
						B	18 min 9 s
						C	18 min 37 s
		$\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	-0.236	25.342	1.241	A	34.4 $mL \cdot kg^{-1} \cdot min^{-1}$
						B	32.1 $mL \cdot kg^{-1} \cdot min^{-1}$
						C	30.4 $mL \cdot kg^{-1} \cdot min^{-1}$
MTA	1 h 48 min	SIM_{MTA}	7.597	-41.037	7.259	A	18 min 24 s
						B	18 min 49 s
						C	19 min 7 s
		$\dot{V}O_2$ max ($mL \cdot kg^{-1} \cdot min^{-1}$)	-1.841	155.472	9.862	A	32.7 $mL \cdot kg^{-1} \cdot min^{-1}$
						B	30.3 $mL \cdot kg^{-1} \cdot min^{-1}$

	C	28.6 mL·kg^{-1}·min^{-1}

Note: OLP=Ordinary Least Products; HART=Hazardous Area Response Team; CBRN=Chemical, Biological, Radiological, Nuclear, MTA=Marauding Terrorist Attack; RIB=Rigid Inflatable Boat; IRU=Incident Response Unit; GTS=Gas-tight Suits. The A, B and C here denote the standards derived from each regression's 90%, 80% and 70% CIs, respectively.

4. Discussion

Specialist emergency responders require adequate physical capability to perform their occupational roles safety and effectively in a variety of hazardous and challenging scenarios. In the present study, we developed novel evidence-based role-specific PES for NARU specialist paramedics. We found that participant performance on the majority of realistic reconstructions of essential, physically demanding tasks were highly correlated with simplified, more easily-replicable simulations of these tasks, with the exception of those replicating water rescue. Occupational tasks were physically demanding and demonstrated moderate-to-very-high correlations with field-expedient gym-based fitness test performance. Collectively, this indicated the components of physical fitness commensurate with job performance and the efficacy of particular simulations and fitness tests for inclusion in physical capability assessments of NARU personnel. In addition, we employed a combination of published and novel methods, first, to derive minimum acceptable standards for each task reconstruction and, second, to produce a range of empirically-linked options for cut-scores on appropriate tests and simulations.

Occupational PES development research follows rigorous best-practice frameworks to ensure resultant role-related tests and standards are empirically linked to job performance and are fair, unbiased and legally defensible (Payne and Harvey 2010; Tipton et al. 2012; Petersen et al. 2016; Reilly et al. 2019). In public-service emergency responders, the predominance of PES research has focused on firefighters (Taylor et al. 2015; Blacker et al. 2015; Siddall et al. 2016; Gumieniak et al. 2018) or law enforcement (Anderson et al. 2001; Morris et al. 2019). While some studies have been completed on ambulance staff (Fischer et al. 2017; Armstrong et al. 2019), the specialist roles of NARU personnel represent a wide-ranging physical and technical capability which warrants specific investigation (Rue et al. 2019). Here, task reconstructions ranged from intense activity bouts that were minutes in duration, to prolonged repetitive tasks of more than an hour. In comparison to the previous NARU demands

analysis, there were few, modest differences in intensity between comparative tasks, but is likely explained by the previous study having conducted tasks in teams and at normal operating speed (rather than at best-effort) (Rue et al. 2019). Our simulations ranged from between ~6-19 minutes in length which is comparable to current occupational testing circuits of ~13-19 minutes in Canadian paramedics (Armstrong et al. 2019) and ~8-12 minutes in UK firefighters (Stevenson et al. 2019). While different occupations are not directly comparable, the range of derived cardiorespiratory fitness standards overlap with standards of non-specialist roles in the UK police (Morris et al. 2019), and the highest we presented (41.1 $mL \cdot kg^{-1} \cdot min^{-1}$; above-ground rescue; HART) is similar to the current minimum PES for UK firefighters (42.3 $mL \cdot kg^{-1} \cdot min^{-1}$; Siddall et al. 2016). The specific correlations between physical fitness tests and task reconstructions we observed are reassuring, as they illustrate a rational mixture of aerobic power, muscular strength, endurance and power that underpin the ability to perform the wide array of tasks, as well as efficacy of field-expedient fitness tests for occupational assessment in this population. This is in agreement with the continued successful use of low-cost, fitness-test batteries within other physically demanding organisations (Gumieniak et al. 2013; Hauschild et al. 2017). We would propose that the observed frequency of aerobic power being highly related to tasks in this study is due to the prolonged nature of the scenarios and the use of performance time as an outcome measure, particularly if greater cardiorespiratory fitness was a surrogate for fewer/shorter recovery periods or better pacing during reconstructions and simulations. This is an important consideration for training in NARU personnel, and is in contrast to more discrete occupational tasks researched in the extant PES literature, where muscular strength and power may appear to more closely contribute to occupational performance.

Measuring or replicating criterion job performance in physically demanding occupations is a critical step in PES development. Research has championed the use of representative simulations of job-tasks (or circuits composed of those tasks) at several stages during the occupational standard-setting process including selection of acceptable job performance (Blacklock et al. 2015; Stevenson et al. 2016), measurement of physical demand (Bilzon et al. 2001; Jamnik et al. 2010a; Taylor et al. 2015; Siddall et al. 2016; Gumieniak et al. 2018) or as end-point assessments in employment testing (Blacker et al. 2015; Burdon and Groeller 2019; Stevenson et al. 2019). We employed similar methods to those

recommended by Blacklock et al (2015), using subject-matter experts to, first, ratify task design, in this case by adjusting previously-designed team-based reconstructions (Rue et al. 2019) into single-person adaptations of those tasks, and second, indicate what constitutes acceptable performance on observation of those tasks being performed. The strengths of scenario-based job simulations are that participants perform the actual physical actions (e.g. movements, technical skill, equipment), and utilise the relevant components of fitness, needed for the criterion job-tasks. For NARU paramedics, the complex and varied nature of potential emergency incidents means that the realistic task reconstructions were appropriate to understand job demands, but would be resource-heavy to use in employment testing. Specifically, the variety of bespoke training areas (e.g. train carriages, scaffolding), supervisory expertise, equipment used, and the geographical spread of ambulance trusts means these reconstructions would likely limit reproducibility, internal consistency and the ability to test frequently. Therefore, we also investigated simplified simulations, constructed with contribution from subject-matter experts and specific criteria to maintain elements of the criterion tasks while being easily-replicable.

With sufficient criterion- and face-validity, assessments using simulations of job-tasks can improve the downstream appeal of PES to incumbents, and subsequent adherence and acceptance of standards (Jamnik et al. 2013; Reilly et al. 2015). Contrary to the relative scarcity of PES research in ambulance workers, a job-task and demands analysis of Canadian paramedics was conducted to successfully develop, validate and implement a job-task simulation involving stretcher/casualty handling as a return-to-work readiness test (Fischer et al. 2017; Armstrong et al. 2019). We observed, with the exception of water-based rescue tasks, high correlations and ratings of similarity between reconstructions and their respective simulations, supporting their overall predictive- and face-validity for assessing occupational capability. The probable reason for lower correlations within water-based tasks is that inter-individual variation in results was due to it relying heavily on technical skill rather than physical capability, and that the simulations did not have the added burden (both external load and to mobility) of the swift water PPE. Anecdotally, participants and observers noted that correctly reading the flow and direction of swift-moving water could greatly aid reaching the casualty while minimising physical effort. Similarly, exiting a conventional swimming pool without PPE did not present the same

technical challenges as using a foot-strop and rope to board a RIB in open water. This may, in part, explain both the lower rating of demand similarity between water-based tasks and simulations, and the illogical finding that boarding the RIB was most highly correlated with estimated $\dot{V}O_2$ max, which probably represents a surrogate of other explanatory factors within a skill-dominant task.

Determining minimum acceptable performance in physically demanding occupations is particularly challenging given its inherent subjectivity, and several accepted subjective- or objective-data-driven methods exist which are preferential for different study designs and objectives (Payne and Harvey 2010). The use of expert judges is well established where, typically, highly-experienced personnel observe different task performances and independently vote on or 'bookmark' what is deemed 'acceptable' between adjacent competencies before collective results are reviewed to reach a consensus (Rogers et al. 2014; Stevenson et al. 2016). Alternatively, experimental-data-driven approaches have been successful, such as aligning physiological demands of tests to field observations, or using normative data to define a standard (e.g. 1 SD beyond mean performance of a representative worker sample) (Bilzon et al. 2001; Jamnik et al. 2010a; Siddall et al. 2016). Any method has potential to introduce bias based on, but not limited to, the experience and/or impartiality of expert raters, the assumed ability of the current workforce and/or measured sample, or the extent to which collected data truly represented safe, reliable and efficient occupational performance (Tipton et al. 2012; Jamnik et al. 2013; Blacklock et al. 2015). In the current study, all performances were conducted at best individual effort in order to obtain correlational relationships between task reconstructions, task simulations and physical fitness tests, acknowledging this is likely faster than normal operational pace. This meant that prospective bookmarking was not possible and applying solely data-driven approaches, which are traditionally intended for normal operational pace, may have been inappropriate. Therefore, we adopted a hybrid approach of triangulating acceptable task performance from subject-matter expert opinion, participant judgement and performance data to attempt to balance potential biases. Consequently, this does not negate the need for the assessing test- and standard-reliability, nor the impact of standards on a novel sample of NARU personnel to assess potential adverse impact.

Two novel aspects of our study were the presentation of options between predictive gym-based tests or job-task simulations for different roles, and production of multiple cut-scores within these individual assessments. These options were derived from the variability present in the relationships between reconstructions and simulations and therefore empirically linked to cohort physical performance. This approach acknowledges the biological variability that occurs in physical performance and, from a practical perspective, provides flexibility for the organisation when implementing and reviewing test standards depending on their latest requirements. Following development of PES, employers often face a number of conflicting practical and organisational factors that influence dissemination of cut-scores and tests (Stevenson et al. 2020). In various industries there may be benefit to having choice of different cut-scores between point-of-entry (pre-training) and incumbent personnel, or different tiers of standards during training or familiarisation (e.g. red-amber-green) (Stevenson et al. 2020). Moreover, there will inherently be error variance in any between-test regression and, particularly when it may impact on decisions of employment, organisations may want to account for this error variance to allow more leniency to the employee. Therefore, the displayed use of compatibility ("confidence") intervals represented a method of producing tiered cut-scores/standards linked to the error variance of each regression analysis.

There are limitations to this work that should mediate interpretation and warrant further work. For a national organisation, NARU is a fairly small group of specialists with approximate personnel numbers of ~750 HART, ~650 MTA and: ~2500 CBRN but where the total will be less than the sum of these values as many personnel cover multiple roles. Our role samples (28 HART, 24 MTA, 23 CBRN) represent 1-5% proportion of the operational NARU roles in the population. While we aimed to examine a representative sample, we acknowledge that those measured may be more or less physically capable than the NARU population, which may have impacted on normative data and perceptions of acceptable standards. The extent to which the study sample was representative of the wider NARU populations can be ascertained through ongoing data capture once the new tests are implemented. Similarly, we paid strict attention to the wording of briefings and questions to frame the notion of minimum acceptable standards to individuals involved, but this study may have placed

proportionally high emphasis on subjective ratings of occupational performance. We have not empirically tested the presence of adverse impact or discrimination to minority groups or to personnel who might typically possess lower occupational capability relative to their peers, which is a vital consideration before the full implementation of any PES (Reilly et al. 2019). Consistent with PES best-practice guidance, we would recommend that this analysis be completed on a sample independent of those used to derive the standards, and in a cohort with large enough sub-groups to provide this evidence. In parallel, we acknowledge that the study did not aim to assess test-retest reliability of the developed simulations and tests. While participants had all completed similar training tasks historically, and were given instructions and demonstrations to familiarise themselves prior to tasks, the time availability of participants and the requirement to reserve specific training locations without disrupting other training meant it would not have been possible to collect well-controlled repeated measures alongside the current study design. The assessment of test reliability is more feasible, and in keeping with normal practice, once the wider NARU workforce have been familiarised to the tests and they can be conducted in their intended workplace settings. Finally, the simulations were produced with the express criteria that they could be administered without use of specialist equipment, locations or the PPE used in the task reconstructions. However, further consideration could have been made to the distribution of external load in simulations, particularly to the feet (to replicate additional mass of footwear), given its specific impact to physiological demand in PPE (Taylor et al. 2012). Better distributed load may have further improved relationships between simulations and reconstructions. This consideration would have been of particular note had the standards been derived by matching metabolic/physiological demand, rather than performance time, which may not be the optimal indicator for "performance" on tasks where safety is of vital concern. Future research should look to ascertain the ability of the developed standards and tests to reliably and correctly discriminate between successful and unsuccessful performers, without adverse impact, in a novel sample of personnel as a part of continual review.

In this study we used a combination of well-established PES research methods and novel approaches to develop role-related standards directly linked to essential job-tasks of NARU personnel.

Given the emerging emphasis on ensuring the health and well-being of paramedic staff (Meadley et al. 2020), we provide a varied and easily-implementable physical capability assessment for NARU personnel to be used to routinely ensure safe and effective work in a vital public-service arena.

Acknowledgements

We are greatly appreciative of the help and logistical support of NARU and all directing staff, instructors and medical support staff at the Defence CBRN Centre, Winterbourne Gunner, UK. Thank you to NARU, National Health Service England, for funding this work. We would also like to express our gratitude to our participants for taking part in this research.

Funding

This work was funded by the National Ambulance Resilience Unit, National Health Service England

References

Anderson GS, Plecas D, Segger T (2001) Police officer physical ability testing – Re-validating a selection criterion. Policing: An International Journal of Police Strategies & Management 24:8–31. https://doi.org/10.1108/13639510110382232

Armstrong DP, Sinden KE, Sendsen J, et al (2019) The Ottawa Paramedic Physical Ability Test: test-retest reliability and analysis of sex-based performance differences. Ergonomics 62:1033–1042. https://doi.org/10.1080/00140139.2019.1618501

Bilzon JLJ, Scarpello EG, Smith CV, et al (2001) Characterization of the metabolic demands of simulated shipboard Royal Navy fire-fighting tasks. Ergonomics 44:766–780. https://doi.org/10.1080/00140130118253

Blacker SD, Rayson MP, Wilkinson DM, et al (2015) Physical employment standards for UK fire and rescue service personnel. Occup Med (Lond). https://doi.org/10.1093/occmed/kqv122

Blacklock RE, Reilly TJ, Spivock M, et al (2015) Standard Establishment Through Scenarios (SETS): A new technique for occupational fitness standards. Work 52:375–383. https://doi.org/10.3233/WOR-152128

Burdon CA, Groeller H (2019) The development of a functional and valid physical employment assessment standard for NSW Mines Rescue Brigadesmen. Work 63:559–569. https://doi.org/10.3233/WOR-192959

Coldwells A, Atkinson G, Reilly T (1994) Sources of variation in back and leg dynamometry. Ergonomics 37:79–86. https://doi.org/10.1080/00140139408963625

Fischer SL, Sinden KE, MacPhee RS (2017) Identifying the critical physical demanding tasks of paramedic work: Towards the development of a physical employment standard. Applied Ergonomics 65:233–239. https://doi.org/10.1016/j.apergo.2017.06.021

Gebhardt DL (2019) Historical perspective on physical employment standards. Work 63:481–494. https://doi.org/10.3233/WOR-192964

Gumieniak RJ, Gledhill N, Jamnik VK (2018) Physical employment standard for Canadian wildland firefighters: examining test-retest reliability and the impact of familiarisation and physical fitness training. Ergonomics 61:1324–1333. https://doi.org/10.1080/00140139.2018.1464213

Gumieniak RJ, Jamnik VK, Gledhill N (2013) Catalog of Canadian fitness screening protocols for public safety occupations that qualify as a bona fide occupational requirement. J Strength Cond Res 27:1168–1173. https://doi.org/10.1519/JSC.0b013e3182667167

Hauschild VD, DeGroot DW, Hall SM, et al (2017) Fitness tests and occupational tasks of military interest: a systematic review of correlations. Occup Environ Med 74:144–153. https://doi.org/10.1136/oemed-2016-103684

Howley ET (2001) Type of activity: resistance, aerobic and leisure versus occupational physical activity. Med Sci Sports Exerc 33:S364-369; discussion S419-420

Jamnik V, Gumienak R, Gledhill N (2013) Developing legally defensible physiological employment standards for prominent physically demanding public safety occupations: a Canadian perspective. Eur J Appl Physiol 113:2447–2457. https://doi.org/10.1007/s00421-013-2603-1

Jamnik VK, Thomas SG, Burr JF, Gledhill N (2010a) Construction, validation, and derivation of performance standards for a fitness test for correctional officer applicants. Appl Physiol Nutr Metab 35:59–70. https://doi.org/10.1139/H09-122

Lentz L, Randall JR, Gross DP, et al (2019) The relationship between physical fitness and occupational injury in emergency responders: A systematic review. Am J Ind Med 62:3–13. https://doi.org/10.1002/ajim.22929

Lockie RG, Dawes JJ, Orr RM, et al (2018) Analysis of the Effects of Sex and Age on Upper- and Lower-Body Power for Law Enforcement Agency Recruits Before Academy Training. The Journal of Strength & Conditioning Research 32:1968–1974. https://doi.org/10.1519/JSC.0000000000002469

Ludbrook J (2012) A primer for biomedical scientists on how to execute model II linear regression analysis. Clin Exp Pharmacol Physiol 39:329–335. https://doi.org/10.1111/j.1440-1681.2011.05643.x

Meadley B, Caldwell J, Perraton L, et al (2020) The health and well-being of paramedics - a professional priority. Occup Med (Lond) 70:149–151. https://doi.org/10.1093/occmed/kqaa039

Morris M, Deery E, Sykes K (2019) Chester treadmill police tests as alternatives to 15-m shuttle running. Occup Med (Lond) 69:133–138. https://doi.org/10.1093/occmed/kqz014

Payne W, Harvey J (2010) A framework for the design and development of physical employment tests and standards. Ergonomics 53:858–871. https://doi.org/10.1080/00140139.2010.489964

Petersen SR, Anderson GS, Tipton MJ, et al (2016) Towards best practice in physical and physiological employment standards. Appl Physiol Nutr Metab 41:S47-62. https://doi.org/10.1139/apnm-2016-0003

Ramsbottom R, Brewer J, Williams C (1988) A progressive shuttle run test to estimate maximal oxygen uptake. British Journal of Sports Medicine 22:141–144. https://doi.org/10.1136/bjsm.22.4.141

Reilly TJ, Gebhardt DL, Billing DC, et al (2015) Development and Implementation of Evidence-Based Physical Employment Standards: Key Challenges in the Military Context. J Strength Cond Res 29 Suppl 11:S28-33. https://doi.org/10.1519/JSC.0000000000001105

Reilly TJ, Neal Baumgartner, Sam Blacker, et al (2019) Combat Integration: Implications for Physical Employment Standards. North Atlantic Treaty Organization

Rogers T, Docherty D, Petersen S (2014) Establishment of performance standards and a cut-score for the Canadian Forces firefighter physical fitness maintenance evaluation (FF PFME). Ergonomics 57:1750–1759. https://doi.org/10.1080/00140139.2014.943680

Rue CA, Rayson MP, Walker EF, et al (2019) A job task analysis to describe the physical demands of specialist paramedic roles in the National Ambulance Resilience Unit (NARU). Work 63:547–557. https://doi.org/10.3233/WOR-192960

Sharp MA, Cohen BS, Boye MW, et al (2017) U.S. Army physical demands study: Identification and validation of the physically demanding tasks of combat arms occupations. Journal of Science and Medicine in Sport 20:S62–S67. https://doi.org/10.1016/j.jsams.2017.09.013

Siddall AG, Stevenson RDM, Turner PFJ, et al (2016) Development of role-related minimum cardiorespiratory fitness standards for firefighters and commanders. Ergonomics 1–9. https://doi.org/10.1080/00140139.2015.1135997

Stevenson RDM, Siddall AG, Turner PFJ, Bilzon JLJ (2016) A Task Analysis Methodology for the Development of Minimum Physical Employment Standards. J Occup Environ Med 58:846–851. https://doi.org/10.1097/JOM.0000000000000812

Stevenson RDM, Siddall AG, Turner PFJ, Bilzon JLJ (2020) Implementation of Physical Employment Standards for Physically Demanding Occupations. Journal of Occupational and Environmental Medicine Publish Ahead of Print: https://doi.org/10.1097/JOM.0000000000001921

Stevenson RDM, Siddall AG, Turner PJF, Bilzon JLJ (2019) Validity and Reliability of Firefighting Simulation Test Performance. J Occup Environ Med 61:479–483. https://doi.org/10.1097/JOM.0000000000001583

Taylor NAS, Fullagar HHK, Sampson JA, et al (2015) Employment Standards for Australian Urban Firefighters: Part 2: The Physiological Demands and the Criterion Tasks. J Occup Environ Med 57:1072–1082. https://doi.org/10.1097/JOM.0000000000000526

Taylor NAS, Lewis MC, Notley SR, Peoples GE (2012) A fractionation of the physiological burden of the personal protective equipment worn by firefighters. Eur J Appl Physiol 112:2913–2921. https://doi.org/10.1007/s00421-011-2267-7

Tipton MJ, Milligan GS, Reilly TJ (2012) Physiological employment standards I. Occupational fitness standards: objectively subjective? Eur J Appl Physiol. https://doi.org/10.1007/s00421-012-2569-4

Wilkinson DM, Blacker SD, Richmond VL, et al (2014) Relationship between the 2.4-km run and multistage shuttle run test performance in military personnel. Mil Med 179:203–207. https://doi.org/10.7205/MILMED-D-13-00291