Reliability of Military-Relevant Tests Designed to Assess Soldier Readiness for Occupational and Combat-Related Duties

Barry A. Spiering, PhD; Leila A. Walker, MS; Nathan R. Hendrickson, MS; Kathleen Simpson, MS; Everett A. Harman, PhD; COL Stephen C. Allison, SP USA (Ret.); Marilyn A. Sharp, MS

ABSTRACT The purpose of this study was to determine the reliability of military-relevant tests designed to assess soldier readiness. Forty-seven soldiers (46 men, 1 woman; 22 ± 3 years; 80.4 ± 11.7 kg) performed each of seven soldier readiness tests on four different occasions over the course of 8 weeks. The soldier readiness tests were: (1) 3.2-km load carriage (LC) time-trial, (2) running long jump (RLJ), (3) one-repetition maximum box lift (1RMBL), (4) 10-minute repetitive box lift and carry (RBLC), (5) simulated victim rescue (VR), (6) mock grenade throw (GT) for accuracy, and (7) simulated combat rushes (CR). Repeated measures analysis of variance revealed significant learning effects for 1RMBL, RBLC, and GT; these tests required two (1RMBL and RBLC) or three (GT) trials to obtain statistically stable values. The intraclass correlation coefficient was 0.78 to 0.89 for all tests. LC, 1RMBL, RBLC, CR, and RLJ all demonstrated standard error of measurement values that were 3% to 5%, whereas VR and GT were 9% and 36%, respectively. In conclusion, the 1RMBL, RBLC, and GT tests required familiarization before a stable value was obtained. The LC, 1RMBL, RBLC, CR, and RLJ tests (and, to a lesser degree, the VR test) demonstrated reasonably acceptable levels of reliability and measurement error, whereas the GT test did not.

INTRODUCTION

Soldiers' occupational and combat-related duties often include physically demanding tasks, such as repetitive lifting, carrying heavy loads over long distances, and swiftly performing battlefield maneuvers within combat environments.[1] Currently, however, the U.S. Army does not routinely assess soldiers' ability to perform such physically demanding, military-relevant tasks. Instead, soldier readiness for occupational and combat-related duties is evaluated using the Army Physical Fitness Test (APFT), which consists of a 2-minute push-up test, a 2-minute sit-up test, and a 3.2-km run. Relying on the APFT as an indicator of soldier readiness is problematic because performance on the APFT poorly predicts performance on common soldiering tasks.[2,3] In recognition of this, the Army is currently evaluating military-relevant methods for assessing soldier readiness (for example, the Army Combat Readiness Test [ACRT]; http://www.armyprt.com/acrt/index.shtml).

Various test batteries have been developed to assess soldier readiness by the Armed Forces of the United States,[3,4] Canada,[5] and the United Kingdom.[6] These test batteries have several common elements. All test batteries assess manual materials handling, such as maximal lifting, repetitive lifting, carrying, combined lifting and carrying, and/or digging, and most test batteries also include some form of load carriage (LC). Logically, these military-relevant tests have been used to assess the effectiveness of various training interventions for improving soldier readiness.[7–13] However, very few studies have evaluated the test–retest reliability of these soldier readiness tests,[5,11,14,15] and the results of these studies are obscured by small sample sizes and/or incomplete statistical analyses. Therefore, the purpose of this study was to assess the test–retest reliability of military-relevant, soldier readiness tests. The objectives of this study were to determine reliability estimates and to help define: (1) the number of trials required to obtain statistically stable results and (2) the precision with which these tests can detect changes in soldier readiness.

United States Army Research Institute of Environmental Medicine, Military Performance Division, 15 Kansas Street, Building 42, Natick, MA 01760.

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Army or the Department of Defense.

METHODS

Experimental Design

Forty-seven soldiers performed each of seven soldier readiness tests on four different occasions over the course of 8 weeks (precise testing schedule provided in Table I). The soldier readiness tests were: (1) 3.2-km LC time-trial while wearing a 33-kg approach load, (2) running long jump (RLJ) while wearing a 20.5-kg fighting load, (3) one-repetition maximum box lift (1RMBL) performed from the ground to the height of 155 cm, (4) 10-minute repetitive box lift and carry (RBLC) task, (5) 50-m simulated victim rescue (VR), (6) 30-m simulated grenade throw (GT) for accuracy, and (7) repeated 30-m combat rushes (CR) while wearing a 20.5-kg fighting load. Repeated measures analysis of variance (ANOVA) was used to identify significant learning effects between trials. Subsequently, the test–retest reliability of each soldier readiness test was assessed via the intraclass correlation coefficient (ICC), standard error of measurement (SEM), and limits of agreement (LOA).

Subjects

An *a priori* sample size analysis indicated that 28 subjects were required to yield statistical power >0.80 assuming an α of 0.05. Therefore, to accommodate missed testing sessions

TABLE I. Testing Schedule

Week	Monday	Tuesday	Wednesday	Thursday	Friday
1	LC	DEXA	RLJ, 1RMBL	RBLC	Rest
2	VR	Rest/Make-Ups	GT, CR	Make-Ups	Rest
3	LC	Rest/Make-Ups	RLJ, 1RMBL	RBLC	Rest
4	VR	Rest/Make-Ups	GT, CR	Make-Ups	Rest
5	LC	Rest/Make-Ups	RLJ, 1RMBL	RBLC	Rest
6	VR	Rest/Make-Ups	GT, CR	Make-Ups	Rest
7	LC	Rest/Make-Ups	RLJ, 1RMBL	RBLC	Rest
8	VR	Rest/Make-Ups	GT, CR	Make-Ups	Rest

1RMBL, one-repetition maximum box lift; CR, repeated combat rushes; DEXA, dual energy x-ray absorptiometry; GT, mock grenade throw for accuracy; LC, 3.2-km load carriage time-trial; RBLC, repetitive box lift and carry; RLJ, running long jump; VR, simulated victim rescue.

and attrition, a total of 47 soldiers (46 men and 1 woman; 22 ± 3 years; 1.75 ± 0.06 m; 80.4 ± 11.7 kg; 18 ± 5% body fat) participated in this study. To be included, soldiers had to have passed their most recent APFT. Soldiers were excluded if they were identified as having a current or previous injury and/or other medical condition that would contraindicate participation, if they recently (within the previous 2 months) began a physical training program designed to increase muscle strength/power, or if they were pregnant. A physician screened all volunteers before participation in the study. Participants were instructed to maintain their normal exercise routine for the duration of the study. Participants were informed of the requirements and potential risks of participation and then voluntarily signed an informed consent document approved by the Institutional Review Board of the U.S. Army Research Institute of Environmental Medicine. The investigators adhered to the policies for protection of human subjects as prescribed in Army Regulation 70-25, and the research was conducted in adherence with the provisions of 45 CFR Part 46.

Procedures

Anthropometrics

Height (cm) and body mass (kg) were measured using a stadiometer and a digital scale, respectively. Body composition was assessed via whole-body scans using fan-beam dual-energy x-ray absorptiometry (DEXA) (Prodigy, GE Medical Systems, Bedford, Massachusetts). Total body estimates of percent fat were determined using manufacturer-described procedures and supplied algorithms (Total Body Analysis, version 3.6, Lunar Corp., Madison, WI).

Load Carriage

The LC test consisted of a 3.2-km time-trial (run and/or walk as fast as possible) on level ground while carrying an approach load (approximately 33 kg), which consisted of weighted vest (15.5 kg), combat boots (2.3 kg), uniform (1.4 kg), helmet (1.4 kg), backpack (9.3 kg), and simulated rifle (3.4 kg). The variable entered into analysis was time-to-completion.

Running Long Jump

Soldiers performed the RLJ while wearing a 20.5-kg fighting load, consisting of a weighted vest (15.5 kg), combat boots (2.3 kg), uniform (1.4 kg), and helmet (1.4 kg). Soldiers began by standing on a line that was 3 m from the take-off point, then running up to the take-off point and leaping forward as far as possible onto a one-inch-thick padded mat. The distance from the take-off point to the landing point closest to the take-off point was measured via a laser distance meter (Disto Plus, Leica Geosystems, Norcross, GA). Participants were given three attempts; the average distance of the three attempts was the variable entered into analysis.

One-Repetition Maximum Box Lift

The 1RMBL test assessed the heaviest box that a soldier could lift onto a 155-cm platform (i.e., the height of a 5-ton Army truck bed). Briefly, the volunteer warmed up by performing 5 repetitions using an unloaded (20.5 kg) box, followed by one repetition each at approximately 50% and approximately 75% of estimated/known 1RM. Three to five subsequent attempts were then used to determine the 1RMBL to the nearest 2 kg, with approximately 3 minutes of rest between attempts. A lift was considered successful if the box was placed onto the platform using proper technique. The variable entered into analysis was the maximal weight lifted onto the platform.

Repetitive Box Lift and Carry

The RBLC test assessed the maximum number of 20.5-kg boxes that a soldier could lift onto a 155-cm platform in 10 minutes. One metal box was positioned at the end of a smooth ramp, which was located 3 m away from and directly in front of the platform. The volunteer lifted the box, carried it 3 m, placed it on the platform, and returned to the starting point for another box. Once the box was lifted from the starting point by the volunteer, a technician at the top of the platform released a second box, which would then slide down the ramp for the subsequent lift. The variable entered into analysis was the number of boxes lifted in 10 minutes.

Victim Rescue

The simulated VR utilized a mannequin (79.5 kg for men, 61.4 kg for the woman) wearing a Modular Lightweight Load-carrying Equipment vest. The mannequin was placed in a seated position on the ground 50 m away from the starting line. The test was performed on a smooth, level

surface with no obstructions. Upon auditory signal from an experimenter, the volunteer ran to the mannequin, grasped the mannequin by the handle on its vest, and dragged it back across the starting line. The variable entered into analysis was the total time required to sprint 50 m and drag the mannequin 50 m back to the starting line.

Grenade Throw

Before initiating the GT test, volunteers warmed up by throwing a tennis ball for approximately 3 minutes. Volunteers began the test in a squatting position, with both feet behind and parallel to a line, with the nonthrowing shoulder pointed toward target. Volunteers then stood up, took one step forward, and threw a mock grenade at a target placed on the ground 30 m away. If more than one step was taken or if the foot crossed over the line, then the throw was not recorded and another throw was completed. For each trial, soldiers were given five practice attempts followed by 1 to 3 minutes rest. Five throws for record were then performed. The distance from where the grenade landed to the center of the target was measured via a laser distance meter. The average distance of the five attempts was the variable entered into analysis.

Combat Rushes

Soldiers performed CR while wearing a 20.5-kg fighting load. Two, one-inch-thick padded mats were positioned 30 m apart. The volunteer began in a prone position on a mat, facing the opposing mat. Following an auditory signal, the volunteer stood up and ran toward the opposing mat; upon reaching the mat, the soldier immediately assumed a prone position and turned to face the direction of the original mat. Five seconds later, there was another auditory signal, which signaled the volunteer to repeat the procedure. This pattern continued until five 30-m rushes were completed. The variable entered into analysis was the total time to complete the task.

Before all testing, volunteers warmed up by jogging in place and performing light stretching as appropriate for the given test. All testing was conducted in a climate-controlled environment whenever possible. The LC test was performed outdoors. Outdoor testing was suspended if the Wet Bulb Globe Temperature (WBGT) was above heat category 2 (31°C WBGT).

Statistical Analyses

Only subjects who completed all four trials for a given test were included in data analysis, resulting in a sample size of 26 to 36 subjects for a given test. The statistical approach used to assess reliability was in accordance with procedures recommended by Atkinson and Nevill.[16] First, repeated-measures ANOVA was used to identify potential learning effects (i.e., significant improvements in performance from one trial to the next). In the event of a significant ($p < 0.05$) F value, a Fisher LSD post hoc test was used to determine pair-wise differences. Next, the reliability of each test was examined. Trials were excluded from this portion of the analysis if significant learning effects occurred for a given test. Furthermore, to standardize the analysis between the various tests, only the first two trials that yielded statistically similar scores were used to assess the reliability of a given test. For example, if there were no significant learning effects for a given test, then Trial 1 and Trial 2 were compared to determine the reliability (i.e., calculate the ICC, SEM, and LOA). Alternatively, if a significant learning effect occurred between Trial 1 and Trial 2, but not between Trial 2 and Trial 3, then Trial 2 and Trial 3 were compared to determine the reliability of the given test.

The reliability of each test was assessed via relative (i.e., ICC) and absolute (i.e., SEM and LOA) reliability statistics [see Ref. 16 for a review of reliability analyses]. ICCs were calculated using a two-way random effects, single-measure reliability model. The SEM was calculated as the square root of the mean squared error term from the ANOVA table. We reported the SEM in absolute units (e.g., cm, s) and relative to the mean value (i.e., percentage of mean). The 95% LOA was calculated as follows: first, the mean scores and residual scores from two consecutive tests were plotted using the methods of Bland and Altman.[17] Then the data were analyzed for the presence of heteroscedasticity by plotting absolute values of individual differences in test–retest scores versus individual means of test–retest scores. Significant Pearson product–moment correlation coefficients were considered indicative of heteroscedastic data (i.e., the random error increased as the mean score increased). If the data were heteroscedastic, then the 95% ratio LOA was calculated as:

$$95\%\ \text{ratio LOA} = [(\text{SD of the difference scores} \div \text{average of the mean values}) \times 1.96] \times 100$$

TABLE II. Performance (Mean ± Standard Deviation) During Repeated Measurements for Individual Soldier Readiness Tests

Test	*n*	Trial 1	Trial 2	Trial 3	Trial 4
LC (s)	26	1430 ± 158	1444 ± 192	1426 ± 229	1430 ± 176
1RMBL (kg)	35	62 ± 10*	66 ± 11	67 ± 11	66 ± 12
RBLC (Number of Boxes)	31	109 ± 9*	114 ± 11	113 ± 12	114 ± 12
VR (s)	34	44.1 ± 8.9	44.0 ± 10.8	42.9 ± 9.4	42.8 ± 9.1
CR (s)	32	61.0 ± 3.7	61.6 ± 3.4	61.7 ± 3.7	61.6 ± 4.0
RLJ (cm)	36	221 ± 34	224 ± 35	223 ± 33	222 ± 35
GT (cm from Target)	34	401 ± 290*	353 ± 252*	311 ± 237	321 ± 255

1RMBL, one-repetition maximum box lift; CR, repeated combat rushes; GT, mock grenade throw for accuracy; LC, 3.2-km load carriage time-trial; RBLC, repetitive box lift and carry; RLJ, running long jump; VR, simulated victim rescue.
*Significantly ($p < 0.05$) different than following trial.

TABLE III. Reliability of Individual Soldier Readiness Tests

Test	*n*	Comparison	ICC(2,1) [95% CI]	SEM (% of Mean)
LC (s)	26	Trial 1 vs. 2	0.81 [0.62–0.91]	77 s (5%)
1RMBL (kg)	35	Trial 2 vs. 3	0.88 [0.78–0.94]	4 kg (5%)
RBLC (Number of Boxes)	31	Trial 2 vs. 3	0.78 [0.59–0.89]	5 boxes (5%)
Simulated VR (s)	34	Trial 1 vs. 2	0.86 [0.73–0.93]	3.8 s (9%)
Repeated CR (s)	32	Trial 1 vs. 2	0.80 [0.63–0.90]	1.6 s (3%)
RLJ (cm)	36	Trial 1 vs. 2	0.89 [0.80–0.94]	11.3 cm (5%)
GT (cm from Target)	34	Trial 3 vs. 4	0.79 [0.62–0.89]	113 cm (36%)

1RMBL, one-repetition maximum box lift; CI, confidence interval; CR, repeated combat rushes; GT, mock grenade throw for accuracy; ICC(2,1), intraclass correlation coefficient, 2-way random-effects model, single measures; LC, 3.2-km load carriage time-trial; RBLC, repetitive box lift and carry; RLJ, running long jump; SEM, standard error of the measurement; VR, simulated victim rescue.

If the data were homoscedastic, then the 95% LOA was reported as:

$$95\%\ \text{LOA} = \text{SD of the difference scores} \times 1.96$$

In practical terms, the LOA indicates that values obtained from two consecutive tests will differ as a result of random measurement error by no more than X units (for LOA) or X% (for ratio LOA) either in the positive or negative direction.[16]

RESULTS

Repeated measures ANOVA revealed significant learning effects for 1RMBL, RBLC, and GT (Table II). 1RMBL and RBLC required two trials to obtain statistically stable values, whereas GT required three trials to obtain statistically stable values. More specifically, 1RMBL increased 8±12% between Trial 1 and Trial 2, RBLC performance improved 4±7% between Trial 1 and Trial 2, and GT accuracy improved 6±35% between Trial 1 and Trial 2 and 6±38% between Trial 2 and Trial 3. All other tests required only one trial to reach statistically stable values.

The ICC(2,1) was 0.78–0.89 for all tests (Table III). LC, 1RMBL, RBLC, CR, and RLJ all produced SEM values that were 3% to 5%, VR produced a SEM value of 9%, and GT produced a SEM of 36% (Table III).

With regards to the LOA analysis, LC, VR, and GT data were heteroscedastic. The 95% ratio LOA for these tests were ±15%, ±24% and ±99%, respectively. The 1RMBL, RBLC, CR, and RLJ data were homoscedastic. The 95% LOA for these tests were ±10 kg, ±15 boxes, ±4.3 s, and ±31.3 cm, respectively.

DISCUSSION

This study revealed three important findings. First, we found that some, but not all, soldier readiness tests required familiarization before statistically stable results were obtained. Specifically, the 1RMBL and RBLC tests required two trials and the GT required three trials. All other tests required only one trial to obtain statistically stable values. Second, all soldier readiness tests demonstrated good "relative reliability" (ICCs = 0.78–0.89). These ICC coefficients indicated that, for a given test, a participant who performed well relative to the other participants during a given trial also tended to perform well during subsequent trials; in other words, the ranking of individuals with respect to the other soldiers tended to stay the same across trials. The third important finding was that, after removing the trials in which significant learning effects occurred, most tests demonstrated very good absolute reliability. Specifically, the LC, 1RMBL, RBLC, CR, and RLJ all produced SEM values that were 3% to 5%. This suggests that these tests can precisely assess changes in soldier readiness following an intervention. The VR demonstrated a "good" SEM value (9%), whereas the GT produced a relatively poor SEM value (36%), indicating that the precision with which the GT test can assess changes in solider readiness is likely to be relatively poor.

There are few previously published results with which to compare our findings. We found that the RBLC test required two trials to obtain statistically stable values. Similarly, previous research also found that repetitive lifting performance improved between the first and second trial,[14,15] with performance stabilizing after the second trial.[14] The ICC of RBLC in the present study (0.78) was slightly lower than the ICC reported in previous investigations (0.94–0.97).[11,14] However, when considering the 95% confidence intervals (CIs) surrounding these estimates (0.59–0.89 in the present study; CIs not reported in the other studies[11,14]), then it seems that our results are reasonably comparable to previous reports. Other research[15] found no significant learning effects for a 15-kg LC task or for maximal box lifting tasks to a height of 145 cm or 170 cm. However, no additional indicators of reliability (e.g., ICC or SEM) were reported in that study.[15] In agreement with those findings,[15] we found that LC performance did not significantly improve from Trial 1 to Trial 2. It is difficult to determine why we found a significant improvement in 1RMBL performance between Trial 1 and Trial 2, whereas others[15] did not, especially because both studies involved sufficiently large samples of soldiers. Ultimately, these comparisons indicate that our findings generally agree with previously published results,[11,14,15] and importantly, this is the first study to systematically evaluate the reliability of a broad range of soldier readiness tests using appropriate statistical analyses.

The LOA analysis used in the present investigation informatively describes the reliability characteristics of the soldier

readiness tests. Such an analysis indicates whether or not the random error increases as the magnitude of the mean value increases (i.e., the presence of heteroscedasticity) and provides a range of values within which the results from any two tests will lie (with 95% probability). With respect to the present study, we found that the LC, VR, and GT tests demonstrated heteroscedasticity, whereas the 1RMBL, RBLC, CR, and RLJ tests did not. In practical terms, this means that soldiers who performed poorly on the LC, VR, and GT tests (i.e., those with the highest mean values) had larger between-trial random variation. For example, the 95% ratio LOA for the LC test was 15%. This implies that a soldier who performs poorly on the LC (e.g., 2000 s) will probably have a larger absolute difference between trials (2000 s × 15% = ±300 s of random error between trials) than a soldier who performs well on the LC (e.g., 1200 s × 15% = ±180 s of random error between trials). Alternatively, for the 1RMBL, RBLC, CR, and RLJ tests, the magnitude of between-trial variation does not depend on the soldier's performance. For example, soldiers demonstrated between-trial random error of ±10 kg during the 1RMBL test, regardless of whether their 1RMBL was 25 kg or 75 kg. In addition to the SEM value, the LOA analysis should assist other researchers in determining whether a given soldier readiness test possesses sufficient precision for use in their investigations.

In addition to describing the test reliability, these data might also help inform policymakers as they seek to improve the assessment of soldier readiness. For instance, in recognition of the poor relationship between APFT performance and soldier readiness,[2,3] the Army is currently in the process of evaluating a new soldier readiness evaluation called the ACRT. The ACRT is a test battery consisting of multiple tests (running speed, VR, carrying, mobility, agility, balance, etc) performed consecutively to assess a soldier's fitness for combat-related duties. Similarly, the Marine Corps Combat Fitness Test (CFT) addresses many of the same physical capabilities as the ACRT and also contains a GT for accuracy. The ACRT and the CFT possess content validity as they assess many components required of soldiers/marines on the battlefield. However, no data currently exist to describe the reliability of these tests. The purpose of this study was not to assess the reliability of the ACRT or the CFT *per se*. However, because the ACRT and the CFT include tests similar to those used in the current investigation (i.e., simulated VR, lifting/carrying, repeated sprints, GTs for accuracy), caution in implementing these tests seems warranted. Specifically, based on the current results, soldiers/marines should be given at least one practice session, on a separate day, before performing the ACRT/CFT for record. This recommendation (i.e., providing familiarization on a separate day) should also apply to collecting normative data to determine appropriate pass/fail standards. If familiarization is not provided, then the standards would likely be set too low because learning effects would not be accounted for. Furthermore, because the ACRT and CFT string together multiple tests, the confounding effects of residual fatigue would likely affect the reliability of a given component test. For example, we found that the VR test required only one trial to obtain statistically stable values and that it yielded a SEM of 9%. However, when administered within the ACRT, the VR test is immediately preceded by several other strenuous tests, which could alter the stability and reliability of the VR results. Future evaluations of the ACRT and CFT should consider the reliability of the entire test battery, as well as the reliability of the individual components tests when administered separately.

In conclusion, we investigated the reliability of seven different tests designed to assess soldier readiness. We found that the 1RMBL, RBLC, and GT tests required familiarization before a stable value was obtained, whereas the LC, VR, CR, and RLJ tests provided stable values after just one test. We also found that the LC, 1RMBL, RBLC, CR, and RLJ tests (and, to a lesser degree, the VR test) demonstrated reasonably acceptable levels of reliability and measurement error, whereas the GT test did not. We feel that the results of this study will provide very practical information for researchers and policymakers who wish to utilize these tests to assess soldier readiness for occupational and combat-related duties.

ACKNOWLEDGMENTS

Technical assistance in developing the testing equipment and procedures was skillfully provided by Mr. Peter Frykman. This study was supported by core funding from the Army Medical Research and Materiel Command.

REFERENCES

1. Sharp MA, Patton JF, Vogel JA: A Database of Physically Demanding Tasks Performed by U.S. Army Soldiers, Technical Report No. T98-12. Natick, MA, U.S. Army Research Institute of Environmental Medicine, 1998. Available at http://www.dtic.mil; accessed January 23, 2012.
2. Knapik JJ, Staab J, Bahrke M, et al: Relationship of Soldier Load Carriage to Physiological Factors, Military Experience and Mood States, Technical Report No. T17-90. Natick, MA, U.S. Army Research Institute of Environmental Medicine, 1990. Available at http://www.dtic.mil; accessed January 23, 2012.
3. Myers DC, Gebhardt DL, Crump CE, Fleishman EA: Validation of the Military Entrance Physical Strength Capacity Test. Alexandria, VA, U.S. Army Research Institute for the Behavioral Sciences, 1984. Available at http://www.dtic.mil; accessed January 23, 2012.
4. Ayoub MM, Jiang BC, Smith JL, Selan JL, McDaniel JW: Establishing a physical criterion for assigning personnel to U.S. Air Force jobs. Am Ind Hyg Assoc J 1987; 48: 464–70.
5. Stevenson JM, Bryant JT, Andrew GM, et al: Development of physical fitness standards for Canadian Armed Forces younger personnel. Can J Sport Sci 1992; 17: 214–21.
6. Rayson M, Holliman D, Belyavin A: Development of physical selection procedures for the British Army. Phase 2: relationship between physical performance tests and criterion tasks. Ergonomics 2000; 43: 73–105.
7. Harman EA, Gutekunst DJ, Frykman PN, et al: Effects of two different eight-week training programs on military physical performance. J Strength Cond Res 2008; 22: 524–34.
8. Hendrickson NR, Sharp MA, Alemany JA, et al: Combined resistance and endurance training improves physical capacity and performance on tactical occupational tasks. Eur J Appl Physiol 2010; 109: 1197–208.

9. Kraemer WJ, Mazzetti SA, Nindl BC, et al: Effect of resistance training on women's strength/power and occupational performances. Med Sci Sports Exerc 2001; 33: 1011–25.
10. Kraemer WJ, Vescovi JD, Volek JS, et al: Effects of concurrent resistance and aerobic training on load-bearing performance and the Army physical fitness test. Mil Med 2004; 169: 994–9.
11. Sharp MA, Harman EA, Boutilier BE, Bovee MW, Kraemer WJ: Progressive resistance training program for improving manual materials handling performance. Work 1993; 3: 62–8.
12. Williams AG, Rayson MP, Jones DA: Effects of basic training on material handling ability and physical fitness of British Army recruits. Ergonomics 1999; 42: 1114–24.
13. Williams AG, Rayson MP, Jones DA: Resistance training and the enhancement of the gains in material-handling ability and physical fitness of British Army recruits during basic training. Ergonomics 2002; 45: 267–79.
14. Pandorf CE, Nindl BC, Montain SJ, et al: Reliability assessment of two militarily relevant occupational physical performance tests. Can J Appl Physiol 2003; 28: 27–37.
15. Rayson M, Holliman D: Physical selection standards for the British Army. Phase 4: predictors of task performance in trained soldiers, DRA/CHS/PHYS/CR95/017, 1-87. Farnborough, UK, United Kingdom Defense Research Agency, 1995. (Available at http://www.mod.uk/DefenceInternet/ContactUs/FindResearchInformation.htm; accessed January 23, 2012.
16. Atkinson G, Nevill AM: Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998; 26: 217–38.
17. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–10.