Research Article   **Open Access**

Retest Repeatability of Motor and Musculoskeletal Fitness Tests for Public Health Monitoring of Adult Populations

Jaana Helena Suni[1]*, Marjo Birgitta Rinne[1] and Jonatan R Ruiz[2]

[1]Urho Kaleva Kekkonen Institute for Health Promotion Research (UKK Institute), Finland
[2]Department of Physical Education and Sport, School of Sport Sciences, University of Granada, Spain

Abstract

Background and purpose: Physical fitness reflects the effects of regular physical activity and is an important prognostic factor of health. Field tests of fitness can be used for public health monitoring, but their accuracy for this purpose needs to be determined. The purpose was to evaluate the adequacy of retest repeatability of nine tests of motor and musculoskeletal fitness against preset criteria.

Methods: One week test-retest measurements with a single tester design were conducted. The participants we volunteers, 25 women and 26 men between ages 22 and 64. The main repeatability estimates of within subject variation were the typical error of measurement and coefficient of variation (CV). For ordinal scale measurements weighted Kappa coefficients were calculated. The CV values 10% or lower and kappa coefficient >0.60 were rated as adequately reliable for public health monitoring. In addition, the systematic chance and percent change in the mean were calculated.

Results: Six tests out of nine showed adequate repeatability (CV ≤ 10% or Kappa coefficient >0.60): agility in of eight run, handgrip strength, lower extremity power in vertical jump, upper body muscular strength and trunk stabilization in modified push-up, muscular endurance in static back extension, and shoulder-neck mobility. Retest learning effect was detected for backwards tandem walk and modified push-up tests. Reliability of two tests was not studied due to a ceiling effect in results towards maximum values. High level of physical activity and fitness of the participants reduced the intra-and inter-individual variation of the test results compared to general population.

Conclusions: Present study provided further knowledge on reliable fitness tests suitable for population monitoring of musculoskeletal functioning and health. These tests may also be used as outcome measures to follow the effects of exercise interventions target to participants with musculoskeletal problems.

Keywords: Physical fitness; Reliability; Surveillance; Adult populations; Musculoskeletal functioning; Health

Abbreviation: CV: Coefficient of Variation

Introduction

Adequate physical activity (PA) and fitness are considered as one of the key factors in current public health promotion. Physical fitness also reflects the effects of regular physical activity. It has been recommended that assessment and monitoring of PA and fitness could be part of a public health strategy [1,2] to deliver interventions to communities likely to increase population PA and fitness levels. Furthermore, data collected from populations should provide evidence based information for health policy planning [1,2].

There is strong evidence that low performance levels in several factors of physical fitness are risk factors for various health problems including the major non-communicable diseases [3-5] musculoskeletal disability related to mobility limitations, [6-8] and increasing evidence for low back pain (LBP) [9-11]. Field-based methods of fitness that show meaningful relationship with health and physical functional ability are needed for promotion of PA and fitness for health. At best these assessments of health-related fitness [12] can be used to monitor the level of fitness in different populations, and to identify those with increased health risks due to inadequate levels of fitness [8,13].

In order to apply fitness tests to large populations the test methods need to be safe, economic and easy to administer under conditions available in ordinary communities [12]. Furthermore, their measurement error (reliability) needs to be established in relation to the measurement purposes [14]. Regarding population based fitness measurements, the testers categorize subjects into different levels of performance (i.e. into fitness classes), make comparisons between individuals and groups, and monitor fitness changes over time [14].

Retest repeatability concerns the consistency of the observed values when the measurement is repeated in same environment, tester and participants. Within-subject variation is the most important type of repeatability measures: the smaller the within-subject variation (i.e. typical measurement error) the better precision of single measurements and better observation of changes [15,16]. In order to correctly categorize individuals into fitness classes, the typical (standard) measurement error of the test needs to be smaller than the average range of applied fitness classes [17]. Correct categorization is a critical issue in targeting interventions to low-fit population groups and individuals with increased risk of diseases or disability, and in epidemiological follow-up studies estimating the predictive effect of fitness level on future morbidity and mortality.

Systematic change in the mean, a non-random change in the measurement value between two test sessions, is an important issue when volunteers perform a series of test trials as part of a monitoring

program [15,16]. A typical example of a systematic change in physical fitness testing results is a learning effect (bias) i.e. the participants perform better in the second test session than the first, because they benefit from the experience of the first test session. To indicate true improvement or deterioration, the change in person's fitness level over time needs to be bigger than the systematic change in the mean. The true change is a critical issue also in epidemiological studies that estimate the predictive effects of fitness changes on future morbidity and mortality [18,19].

The purpose of this study was to evaluate the retest repeatability of a set of health-related fitness tests which have potential to be used in public health monitoring to predict changes in future health and musculoskeletal functioning.

Materials and Methods

Subjects

Volunteers from two departments of the University of Tampere,

	MEN n=26	WOMEN n=25	ALL n=51
Civil status	%	%	%
Single	23	24	23
cohabiting	77	76	77
Education			
low (vocational)	39	8	24
mid (high school)	42	16	30
high (university)	19	76	47
Smoking			
never	73	92	82
quit over a year ago	19	4	12
occasionally	4	4	4
daily	4	0	2
Physical strain at work*			
light	65	76	71
moderate	35	20	27
not working	0	4	2
Perceived health†			
average	31	20	25
good	50	72	61
very good	19	8	14
Musculoskeletal problems			
none	77	28	53
back problems	11	28	19
neck problems	4	20	12
lower extremity problems	4	8	6
multiple problems	0	4	2
some other	4	12	8
Overall level of physical activity‡			
only light activity 2-7 times/week	12	8	10
brisk exercise 1-2 times/ week	23	44	33
brisk exercise at least 3 times/week	65	48	57
Self-reported fitness level (compared to own age group)			
markedly worse	0	4	2
somewhat worse	4	0	2
similar	39	56	47
somewhat better	42	32	37
markedly better	15	8	12

*none of the participants was engaged in strenuous physical work
†none of the participant reported poor or very poor health
‡none of participant reported being physically inactive

Table 1: Background characteristics of the study participants.

Finnish Railway Company Ltd, and a private gear manufacturing enterprise were recruited by sending an invitation letter with detailed information on the study by email to persons in supervising positions, who then further delivered the message. Altogether 51 subjects participated in the study, 25 women and 26 men between ages 22 to 64. The mean age of the men was 42.5 (SD 13.4) and of the women 39.4 (SD 9.8). Background characteristics of the participants are given in Table 1.

Procedures

The subjects were asked to contact the health promotion research institute by phone or email to make appointments for two measurement sessions approximately a week apart. The mean number of days between the two measurement sessions was 7.0 (SD 2.1, range from 2 to 15), 82% (n=40) managed to have a retest within 5-9 days.

A pre-testing health screening was conducted according to the safety model of the Health-Related Fitness Test Battery for Middle-aged Adults [20]. It included a modified version of the Physical Activity Readiness Questionnaire (MPAR-Q) [21], and questions on perceived health status, overall level of physical activity [22] (including weekly frequency and intensity), and type of exercise. One experienced fitness tester was responsible for conducting all the fitness measurements.

Assessment of health-related fitness

Nine fitness tests were individually performed by each subject in a pre-set sequence with standard methods [17,23,24]: three motor tests measuring balance and agility, and six musculoskeletal tests measuring shoulder-neck flexibility, muscular strength of upper and lower extremities, and endurance of the trunk muscles. The fitness items with description of their purpose, exclusion criteria, test performance and instructions, and scoring are presented in Appendix.

Statistical analyses

The mean with standard deviation (SD) or the median, and the minimum and maximum of test-retest results in women and men are presented as descriptive statistics. The estimates of repeatability for interval scale measurements were calculated as suggested by Hopkins [15]: (1) typical (standard) error of measurement (s), indicating within-subject variation, was calculated as the standard deviation of test-retests difference (SD_{diff}) divided by the square root of two ($s=SD_{diff}/\sqrt{2}$). To compare the repeatability of different tests the typical error was presented also as the relative measurement error i.e. (2) coefficient of variation (CV): typical error divided by the mean of two tests (CV=s/mean($test_i$+$retest_i$)/2). The (3) systematic chance in the mean with 95% confidence interval was calculated using paired samples t-test in SPSS 17.0 for Windows software (SPSS Inc., Chicago, IL). The mean change between the two measurement sessions was considered statistically significant if the 95% confidence interval (CI) did not include value zero. The (4) percent change in the mean performance between the first and second test session was also calculated. In addition, (5) the Bland and Altman plots (data not shown) were screened for heteroscedasticity [14]. For ordinal scale measurements (6) weighted Kappa coefficient was used to estimate the reproducibility between two test sessions.

The rationale for the criteria on adequacy of repeatability in the present study

Currently there are no consensus statements or standards for acceptable measurement precision for monitoring physical fitness in population level. Other risk factors of non-communicable diseases such as blood pressure, cholesterol and length of waist circumference have been systematically monitored among large populations in several

countries around the world. According to data from Australia [25] the CV for intra-individual variation of waist circumference was less than 1%, and less than 5% for blood pressure. The intra-individual variation in total cholesterol measurements is 7% and additional analytical errors of laboratory personnel 6% (personal information from a chief biochemist of a university hospital). Due to this fairly large total error (13%), it has been recommended that measurements of cholesterol level should be taken more than one time before making decisions on medication.

We applied the above knowledge on the aforementioned CVs while making our definitions on the adequacy of retest reliability of fitness tests aimed for public health monitoring. The CV values of intra-individual variation (i.e. typical error) were rated as follows: less than or equal to 5% as highly reliable, values between 6 and 10% adequately reliable, and those 11% or more not reliable for public health monitoring. Altman [26] has rated the Kappa coefficients as follows: 0.81 to 1.00=very good, 0.61 to 0.80 = good, 0.41 to 0.60 = moderate, 0.21 to 0.40=fair and ≤ 0.20=poor. In the present study weighted Kappa estimates higher than 0.60 were considered adequately repeatable for the purpose of public health monitoring.

Results

Descriptive results of test-retests measurements in women and men are presented in Table 2. In general the men had somewhat better fitness in all tests except the static back extension in which women had better endurance strength. In the shoulder-neck test women had better result for right side than the men: in the first and second session 13 women of 25 had maximum score of 3 (no movement restriction), respective figures for men were six and seven out of 26. There was a strong ceiling effect towards maximum values in one motor fitness test (balance in one-leg stand) and one musculoskeletal test (muscular endurance in dynamic sit-up). Due to this repeatability analyses were not conducted for these two tests items.

The main result of the retest repeatability analyses for interval scale measurements are presented in Table 3. Regarding *the tests of motor fitness*, the smallest within-subject variation as the percent typical error (i.e. CV) was found in for running the figure-of-eight test (2%) with standard deviation of the mean difference (SD_{diff}) being 0.22 s. Percent change in the mean between the two test sessions was less than 1%, and there was no sign of systematic bias. Repeatability of the backwards tandem walk test indicated more intra-individual variation (SD_{diff}=3.2s) the CV of 15% being the highest of all of the tests analysed (Table 3). Furthermore, the performance significantly improved from test to retest.

Of *the musculo skeletal tests*, handgrip strength showed small intra-individual variation (SD_{diff}=2.3 kg) the CV being the second lowest

HRF COMPONENT			Women				Men			
Fitness test										
MOTOR FITNESS										
One-leg stand (sec.)	N	Median		Min	Max	N	Median		Min	Max
1st measurement	25	60.0		60	60	26*	60.0		35	60
2nd measurement	25†	60.0		30	60	26‡	60.0		28	60
Backwards tandem walk (sec.)	N	Mean	sd	Min	Max	N	Mean	sd	Min	Max
1st measurement	25	17.4	5.4	12	37	26	14.4	4.0	9	28
2nd measurement	25	15.4	3.4	11	23	26	13.0	2.8	9	21
Running figure-of-eight (sec.)	N	Mean	sd	Min	Max	N	Mean	sd	Min	Max
1st measurement	25	6.9	0.5	6.2	8.3	26	6.1	0.6	5.2	7.2
2nd measurement	24	7.0	0.6	6.1	8.3	25	5.9	0.5	5.2	7.2
MUSCULOSKELETAL FITNESS										
Hand-grip strength (kg)	N	Mean	sd	Min	Max	N	Mean	sd	Min	Max
1st measurement	25	33.9	5.2	23.0	45.0	26	51.1	6.3	38.5	60.5
2nd measurement	25	35.3	5.3	25.5	49.0	26	52.6	6.9	34.5	63.5
Jump-and-reach (cm)	N	Mean	sd	Min	Max	N	Mean	sd	Min	Max
1st measurement	25	32.7	5.2	21.0	42.0	26	48.3	10.3	25.0	74.0
2nd measurement	24	31.5	6.6	18.0	43.0	25	49.0	8.8	29.0	68.0
Dynamic sit-up (repetitions)	N	Median		Min	Max	N	Median		Min	Max
1st measurement	25†	15.0		4	15	26	15.0		15	15
2nd measurement	24†	15.0		5	15	26	15.0		15	15
Modified push-up (repetitions)	N	Mean	sd	Min	Max	N	Mean	sd	Min	Max
1st measurement	25	12.4	2.6	8	17	26	15.3	2.7	10	21
2nd measurement	24	13.4	3.1	8	18	25	16.6	3.1	12	24
Static back extension (sec.)	N	Mean	sd	Min	Max	N	Mean	sd	Min	Max
1st measurement	25	129.8	35.5	60	180	26	124.2	31.5	64	180
2nd measurement	24	132.9	34.7	73	180	26	122.6	29.3	85	180
Shoulder-neck mobility (points)		Score=1	Score=2	Score=3	Median		Score=1	Score=2	Score=3	Median
1st measurement, right	25	n=2	n=10	n=13	3	26	n=5	n=15	n=6	2
2nd measurement, right	25	n=3	n=9	n=13	3	26	n=9	n=10	n=7	2
1st measurement, left	25	n=2	n=12	n=11	2	26	n=7	n=13	n=6	2
2nd measurement, left	25	n=3	n=12	n=10	2	26	n=8	n=11	n=7	2

*four male participants did not complete the test with maximum result
†one female participant did not complete the test with maximum result
‡two male participants did not complete the test with maximum result

Table 2: Descriptive results of test-retest measurements of health-related fitness (HRF) in women and men.

FITNESS COMPONENT		Mean of	Typical error		Change in the mean:		
Fitness test item	n	two test	s	CV(%)	mean (95% CI)		%†
MOTOR							
One-leg stand (s)	51	58.8	na	na	na		na
Backwards tandem walk (s)	51	15.1	2.2	15	-1.7	(-2.6 to -0.8)	10
Running figure-of-eight (s)	49	6.47	0.26	2	-0.03	(-0.04 to 0.09)	0.4
MUSCULOSKELETAL							
Hand-grip strength (kg)	51	43.4	1.9	4	1.4	(0.6 to 2.2)	3
Jump-and-reach (cm)	49	40.6	2.4	6	-0.4	(-1.4 to 0.6)	1
Dynamic sit-up (repetitions)	50	14.9	na	na	na		na
Modified push-up (repetitions)	49	14.5	1.1	8	1.1	(0.7 to 1.6)	8
Static back extension (s)	50	126.7	13.1	10	1.6	(-3.7 to 6.9)	1

s=typical error as standard deviation of the intra-individual test-retests difference divided by the square root of two (s=$sd_{diff}/\sqrt{2}$).
CV=coefficient of variation as typical error divided by the mean of two tests =$s/((test_i+retest_i)/2)$
95% CI= 95% confidence interval
†percent change in the mean= change in the mean divided by the mean of first test =$mean(test_i-retest_i)/mean(test_i)$
na=not available due to strong positive ceiling effect towards maximum value.

Table 3: One week test-retest repeatability results on field tests of health-related fitness.

(4%) of all tests. However the 95% CI of the change in the mean indicate a small, statistically significant improvement in handgrip strength from first test session to retest session (Table 3). Lower extremity power in the jump-and-reach test also had fairly small intra-individual variation (SD_{diff}=3.4cm) and CV 6%, and a very low 1% change in the mean with no indication of systematic bias. The CV of intra-individual variation (SD_{diff}=1.6rep.) in modified push-up test was 8%, percentage change in the mean being also 8. The retest performance improved on average, 1.1 repetitions, which was statistically significant (Table 3). Endurance time in static back extension test (SD_{diff}=18.6s) had CV of 10%, however the mean change value was small with percent change of 1% (Table 3).

In the shoulder-neck flexibility test, the result was based on visual observation of the tester with ordinal scale criteria for three categories (Appendix). The weighted Kappa coefficient for right side was good (0.80) and very good (0.85) for left side. The number of discordant pairs from test to retest for the right side was eight (2 improved, 6 deteriorated) and six for left side (2 improved, 4 deteriorated). None of the ratings changed from 1 to 3 or vice versa.

Discussion

We studied different aspects of reliability of nine field tests of motor and musculoskeletal fitness with potential for public health monitoring among working aged populations. The following six tests showed adequate repeatability according to preset criteria (CV ≤ 10%, Kappa coefficient >0.60, see methods) on intra-individual variation within one week's test-retest sessions for population monitoring: agility in figure-of-eight run, handgrip strength, lower extremity power in vertical Jump (jump-and-reach), upper body muscular strength and trunk stabilization in modified push-up, muscular endurance in static back extension, and shoulder-neck mobility. Backwards tandem walk test had unacceptably high intra-individual variation and there was a systematic learning effect. Tests of one-leg balance and dynamic sit-ups lacked discriminatory power (Table 2) and were excluded from the repeatability analyses.

Repeatability of the motor fitness tests

The *running figure-of-eight* test had the lowest intra individual variation of all the tests (CV 2%), and very small change in the mean. Absolute measures of repeatability for the test were first introduced

by Vartiainen et al. [24] in young healthy men. They reported the typical error to be 0.14s which is similar to the present finding of 0.16s in healthy middle-aged men and women. Among 50 young women (athletes and non-athletes) [27] zigzag run test had higher repeatability correlations (intra class correlation coefficient i.e. ICC was 0.97) than any of the five studied hop tests (range of ICC from 0.84 to 0.94, standard error of measurement 0.26s) which may indicate easier customizing to running compared to hopping. The *running figure-of-eight* test was originally designed for clinical use to monitor patients after knee trauma operations [28]. The test requires both agility and speed of movement (power). More recently, a high performance level in the test was associated with high quality of life in elderly women [29]. Furthermore, the test result seems to differentiate patients with head injury (contusion) from healthy control persons [30]. Thus, figure of eight test seems to offer a reliable and meaningful health-related test mode that is feasible from early adulthood to old age.

The repeatability of the of the *backwards tandem walk* test was poor (CV 15%). Originally the test used as an indicator of dynamic balance in elderly women [31], and the authors reported high correlation (r=0.94) between the repeated measures one week apart. High correlation (ICC=0.85) between two tests and the 95% limits of agreement in the Bland and Altman plots were also reported by Rinne et al. [23]. Their conclusion of the repeatability was "reasonable." The most recent findings by Vartiainen et al. [24] agree quite well with the present findings: ICC 0.71, SD_{diff} 2.2s, and typical error 1.6s with corresponding CV of 12% (the latter calculated by the present authors). Among the elderly, both poor static and dynamic balance are strong predictor of mobility disability [8,32] and falls [33].

Repeatability of the musculoskeletal fitness tests

In the present study measurements of *hand-grip strength* were highly repeatable (CV 4%) among working aged men and women regardless of a small learning effect. The former studies have also reported high repeatability correlations [34,35], however, there is considerable variation and insufficient reporting on the measurement protocols, values recorded, and results used (mean value, from one, two or three attempts, with either hand or the dominant hand alone) [35]. For instance, body, wrist and handle positioning have effect on both strength values and test repeatability [35]. The latter ones were all

controlled in the present study (see Appendix). Low grip strength has been associated with a greater likelihood of premature mortality and the development of disability in middle-aged and elderly populations [36,37], and it is measured to detect sarcopenia [34].

The intra-individual variation in test-retest sessions of jump-and-reach test (i.e. vertical jump) was adequate (CV 6%) with no systematic bias and very low change in the mean (1%). The finding is line with a former study, which reported standard error of 3.0 cm [17]. According to the review by Hopkins et al. [16] the CV of tape measures of vertical jump height among athletes varied between 3.8-4.8%. Athletes are likely to be measured regularly however our results with more "novice" jumpers are well in line with these former results. In the present study protocol, a practice trial was performed before the two test trials in both test sessions. This is important, while the CV between subsequent trials during same measurement session [16] is much smaller between second and third trial than (0.2%) than the first and second trial (1.2%). Vertical jump requires ability to activate fast type of motor units in a short period of time i.e. muscular power. The importance of power of lower extremities has been emphasized in relation to mobility disability and ultimately risk of falls in elderly populations, especially in women [38-40]. While muscular power starts to decline after 40 year of age which is earlier than maximal muscular strength or endurance [41], population monitoring should start before this age to enhance the importance of muscular power type of training for maintenance of mobility functioning at older age.

The *modified push-up* test results showed acceptable within- subject variation (CV 8%), however there was a statistically significant mean improvement in test-retest results of 1.1 repetitions (95% CI 0.7 to 1.6) indicating a learning effect. The participants were allowed to practice the performance technique once during a push-up cycle in order to not to cause fatigue before the actual testing. In a previous investigation with a less selected study population, the test-retest leaning effect was even bigger (mean difference 3.0; 95% CI from 2.1 to 3.9) [17]. A practice test session before actual testing would probably help to overcome the bias of learning. We were not able to find any studies reporting absolute measures of repeatability on conventional push-up tests to compare the results with the modified test. Low fitness in modified push-up test has been associated with poor perceived health, and low back dysfunction and pain among middle aged [42]. Recently, an increased risk for low back pain was reported in previously healthy conscripts with a poor fitness level in trunk muscular endurance and aerobic performance [11]. The strongest risk factor at entry was poor fitness in both back-lift and conventional push-up tests, i.e., co-impairment (hazard ratio 2.8; 95% confidence interval 1.4-5.9).

The intra-individual variation in *static back (trunk) extension test* was barely acceptable (CV 10%). The test is a modification of the Sørensen back endurance test. A review by Essendrop et al. [43] concluded that there is lack of repeatability statistics beside Pearson correlations of the Sørensen test. Keller et al. [44] reported CV of 21% and stated that the test-retest variation was too high for use in follow-up studies. The results of the present study were somewhat better (CV 10%), possibly due to afore mentioned selection of the subjects. It is well accepted that motivation is strongly related to the back endurance test performance due to uncomfortable static position and a long testing time. Despite of the poor general repeatability of the static trunk extension tests, there is evidence that poor endurance predicts the incidence of LBP [11,45-47].

In the *shoulder-neck mobility* test the weighted Kappa coefficients were well acceptable (≥0.80) for population monitoring. The results were somewhat better than in a former study where weighted Kappa for right side was 0.60 and for left 0.61 [17]. This test could be a practical tool to screen out subjects with stooped upper body posture (see Appendix) commonly related to high amount of sitting and computer work in children, adolescents and working age population. In time, the stooped posture leads to restricted shoulder joint mobility and changes in vertebral level mobility of the lower cervical and upper thoracic spine.

Limitations of the study

The main limitation of the study was the homogeneity of the participants in terms of high level of physical activity and fitness which became restrictive for evaluations of repeatability of two tests (one-leg stand, dynamic sit-up) and probably reduced the intra- and inter-individual variation in all test results. The results of the present study are likely to overestimate the level of repeatability in situations when the studied fitness tests are applied to general population.

Strengths of the study

Reasonable precision for estimates of reliability requires approximately 50 study participants [15], the present study included 51 subjects. Furthermore, only one experienced fitness tester conducted the measurements, which emphasizes the role of intra-individual variation as the source of error [15]. The selected statistical methods and outcome measures agree with expert recommendations [14-16,26]. We included Bland Altman plots to screen out possible heteroscedasticity [14], but decided not publish them in the present paper while no such trends were detected. We chose not the report the 95% limits of agreement, while they are considered too wide (2.8 time typical errors) to detect important trustworthy changes [15].

Implications

Musculoskeletal disorders, constitute major public health problems, and are the most common causes of work disability and consequent absence from work [48]. The results of the present study provide further knowledge on reliable measurement tools for population monitoring of musculoskeletal functioning and health. Physiotherapists could also use the tests as outcome measures of activity limitations in exercise intervention targeted to improve physical functioning of working aged individuals with musculoskeletal problems.

Acknowledgement

References

1. Bull FC, Bauman AE (2008) Consistent risk factor monitoring systems underpins good public health practice. Prev Med 47: 154-155.

2. Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, et al. (2012) Global physical activity levels: surveillance progress, pitfalls, and prospects. Lancet 380: 247-257.

3. Kodama S, Saito K, Tanaka S, Maki M, Yachi Y, et al. (2009) Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: a meta-analysis. JAMA 301: 2024-2035.

4. Fogelholm M (2010) Physical activity, fitness and fatness: relations to mortality, morbidity and disease risk factors. A systematic review. Obes Rev 11: 202-221.

5. Sawada SS, Lee IM, Naito H, Noguchi J, Tsukamoto K, et al. (2010) Long-term trends in cardiorespiratory fitness and the incidence of type 2 diabetes. Diabetes Care 33: 1353-1357.

6. Guralnik JM, Ferrucci L, Pieper CF, Leveille SG, Markides KS, et al. (2000) Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. J Gerontol Biol Sci Med Sci 55: M221-M231.

7. Stratford PW, Kennedy DM, Woodhouse LJ (2006) Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 86: 1489-1496.

8. Husu P, Suni J, Pasanen M, Miilunpalo S (2007) Health-related fitness tests as predictors of difficulties in long-distance walking among high-functioning older adults. Aging Clin Exp Res 19: 444-450.

9. Husu P, Suni J (2012) Predictive validity of health-related fitness tests on back pain and related disability: a 6-year follow-up study among high-functioning older adults. J Phys Act Health 9: 249-258.

10. Heneweer H, Picavet HS, Staes F, Kiers H, Vanhees L (2012) Physical fitness, rather than self-reported physical activities, is more strongly associated with low back pain: evidence from a working population. Eur Spine J 21: 1265-1272.

11. Taanila HP, Suni JH, Pihlajamäki HK, Mattila VM, Ohrankämmen O, et al. (2012) Predictors of low back pain in physically active conscripts with special emphasis on muscular fitness. Spine J 12: 737-748.

12. Skinner J, Oja P (1994) Laboratory and field tests for assessing health-related fitness: Physical Activity, Fitness, and Health. : Human Kinetics, Champaign, IL.

13. American College of Sports Medicine (2010) ACSM's Health-Related Physical Fitness Assessment Manual. Lippincott Williams & Willkins, Philadelphia, PA.

14. Atkinson G, Nevill AM (1998) Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 26: 217-238.

15. Hopkins WG (2000) Measures of reliability in sports medicine and science. Sports Med 30: 1-15.

16. Hopkins WG, Schabort EJ, Hawley JA (2001) Reliability of power in physical performance tests. Sports Med 31: 211-234.

17. Suni JH, Oja P, Laukkanen RT, Miilunpalo SI, Pasanen ME, et al. (1996) Health-related fitness test battery for adults: aspects of reliability. Arch Phys Med Rehabil 77: 399-405.

18. Williams PT (2003) The illusion of improved physical fitness and reduced mortality. Med Sci Sports Exerc 35: 736-740.

19. Jackson AS, Kampert JB, Barlow CE, Morrow JR Jr, Church TS, et al. (2004) Longitudinal changes in cardiorespiratory fitness: measurement error or true change? Med Sci Sports Exerc 36: 1175-1180.

20. Suni JH, Miilunpalo SI, Asikainen TM, Laukkanen RT, Oja P, et al. (1998) Safety and feasibility of a health-related fitness test battery for adults. Phys Ther 78: 134-148.

21. Thomas S, Reading J, Shephard RJ (1992) Revision of the Physical Activity Readiness Questionnaire (PAR-Q). Can J Sport Sci 17: 338-345.

22. Fogelholm M, Malmberg J, Suni J, Santtila M, Kyröläinen H, et al. (2006) International Physical Activity Questionnaire: Validity against fitness. Med Sci Sports Exerc 38: 753-760.

23. Rinne MB, Pasanen ME, Miilunpalo SI, Oja P (2001) Test-retest reproducibility and inter-rater reliability of a motor skill test battery for adults. Int J Sports Med 22: 192-200.

24. Vartiainen MV, Rinne MB, Lehto TM, Pasanen ME, Sarajuuri JM, et al. (2006) The test-retest reliability of motor performance measures after traumatic brain injury. Adv Physiother. 8: 50-59.

25. Dhaliwal SS, Welborn TA (2009) Measurement error and ethnic comparisons of measures of abdominal obesity. Prev Med 49: 148-152.

26. Altman DG (1991) Practical Statistics for Medical Research. Chapman and Hal, London.

27. Ortiz A, Olson SL, Roddey TS, Morales J (2005) Reliability of selected physical performance tests in young adult women. J Strength Cond Res 19: 39-44.

28. Tegner Y, Lysholm J, Lysholm M, Gillquist J (1986) A performance test to monitor rehabilitation and evaluate anterior cruciate ligament injuries. Am J Sports Med 14: 156-159.

29. Karinkanta S, Heinonen A, Sievanen H, Uusi-Rasi K, Kannus P (2005) Factors predicting dynamic balance and quality of life in home-dwelling elderly women. Gerontology 51: 116-121.

30. Rinne MB, Pasanen ME, Vartiainen MV, Lehto TM, Sarajuuri JM, et al. (2006) Motor performance in physically well-recovered men with traumatic brain injury. J Rehabil Med 38: 224-229.

31. Nelson HD, Nevitt MC, Scott JC, Stone KL, Cummings SR (1994) Smoking, alcohol, and neuromuscular and physical function of older women. Study of Osteoporotic Fractures Research Group. JAMA 272: 1825-1831.

32. Rantanen T, Guralnik JM, Ferrucci L, Leveille S, Fried LP (1999) Coimpairments: strength and balance as predictors of severe walking disability. J Gerontol A Biol Sci Med Sci 54: M172-176.

33. Karinkanta S, Piirtola M, Sievänen H, Uusi-Rasi K, Kannus P (2010) Physical therapy approaches to reduce fall and fracture risk among older adults. Nat Rev Endocrinol 6: 396-407.

34. Roberts HC, Denison HJ, Martin HJ, Patel HP, Syddall H, et al. (2011) A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach. Age Ageing 40: 423-429.

35. Fong PW, Ng GY (2001) Effect of wrist positioning on the repeatability and strength of power grip. Am J Occup Ther 55: 212-216.

36. Rantanen T, Guralnik JM, Foley D, Masaki K, Leveille S, et al. (1999) Midlife hand grip strength as a predictor of old age disability. JAMA 281: 558-560.

37. Bohannon RW (2008) Hand-grip dynamometry predicts future outcomes in aging adults. J Geriatr Phys Ther 31: 3-10.

38. Marsh AP, Miller ME, Saikin AM, Rejeski WJ, Hu N, et al. (2006) Lower extremity strength and power are associated with 400-meter walk time in older adults: The CHIANTI Study. J Gerontol A Biol Sci Med Sci 61: 1186-1193.

39. Larsen AH, Sørensen H, Puggaard L, Aagaard P (2009) Biomechanical determinants of maximal stair climbing capacity in healthy elderly women. Scand J Med Sci Sports 19: 678-686.

40. Raj IS, Bird SR, Shield AJ (2010) Aging and the force-velocity relationship of muscles. Exp Gerontol 45: 81-90.

41. Deschenes MR (2004) Effects of aging on muscle fibre type and size. Sports Med 34: 809-824.

42. Suni JH, Oja P, Miilunpalo SI, Pasanen ME, Vuori IM, et al. (1998) Health-related fitness test battery for adults: associations with perceived health, mobility, and back function and symptoms. Arch Phys Med Rehabil 79: 559-569.

43. Essendrop M, Maul I, Läubli T, Riihimäki H, Schibye B (2002) Measures of low back function: a review of reproducibility studies. Clin Biomech (Bristol, Avon) 17: 235-249.

44. Keller A, Hellesnes J, Brox JI (2001) Reliability of the isokinetic trunk extensor test, Biering-Sørensen test, and Astrand bicycle test: assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. Spine (Phila Pa 1976) 26: 771-777.

45. Hamberg-van Reenen HH, Ariëns GA, Blatter BM, Twisk JW, van Mechelen W, et al. (2006) Physical capacity in relation to low back, neck, or shoulder pain in a working population. Occup Environ Med 63: 371-377.

46. Strøyer J, Jensen LD (2008) The role of physical fitness as risk indicator of increased low back pain intensity among people working with physically and mentally disabled persons: a 30-month prospective study. Spine 33: 546-554.

47. Alaranta H, Luoto S, Heliövaara M, Hurri H (1995) Static back endurance and the risk of low-back pain. Clin Biomech (Bristol, Avon) 10: 323-324.

48. Bevan S, Quadrello T, McGee R, Mahdon M, Vovrovsky A, et al. (2009) Fit For Work-Musculoskeletal disorders in the European workforce.