Reliability of the JobFit System Pre-Employment Functional Assessment Tool

Jennifer Legge[a,b,*] and Robin Burgess-Limerick[b,c]
[a]*JobFit Systems International Pty Ltd, PO Box 8740, Mt Pleasant, QLD 4740, Australia*
[b]*School of Human Movement Studies, University of Queensland, St Lucia, QLD 4067, Australia*
[c]*Burgess-Limerick & Associates, 12 Ardoyne Road, Corinda, QLD 4075, Australia*

Received 21 January 2006
Accepted 14 March 2006

Abstract. Functional capacity testing in the pre-employment or post-offer phase of recruitment is increasing in popularity as a preventative tool for controlling sprains and strains in the workplace. The purpose of this study is to determine the reliability of the JobFit System Pre-Employment Functional Assessment (PEFA) as a whole, or in parts, as a precursor for a validity study investigating the relationship between PEFA results and workplace injury rates and severity.

A group of 28 healthy male coal mine employees were videotaped whilst they participated in a generic JobFit System Pre-Employment Functional Assessment (PEFA) including tests of aerobic physical fitness, balance, postural tolerances and material handling tolerances. Twenty participants performed a second trial. The test component scores and overall PEFA scores were compared between trials (test-retest, intra-rater) and assessors (inter-rater) to determine their reliability expressed in terms of ICC.

Using an ICC score of > 0.75 as good and > 0.90 as excellent, in conjunction with percentage agreement a good to excellent reliability rating was allocated to the overall PEFA score, floor to bench lift, bench to overhead lift, bilateral carry and climbing. A moderate to good rating was recorded for bench to shoulder lifts, reaching forward, reaching overhead and stooping. A poor to moderate rating was recorded for squatting, balance and fitness tests. Test-retest scores were typically lower than intra-tester and inter-tester scores. ICC scores should be interpreted with consideration of their limitations and in conjunction with the actual test results.

Keywords: Reliability, pre-employment functional assessment, work-related assessment, functional capacity evaluation, pre-placement assessment

1. Introduction

Work-related musculoskeletal injuries cost companies millions of dollars every year in the form of reduced productivity, replacement wages, medical costs, lump sum payments and performance-based workers compensation premiums. According to the Australian National Occupational Health and Safety Commission there were 138,810 new compensation claims in 2001/02 over half of which (54%) were due to sprains and strains. 41% of all cases were due to body stressing (manual handling) with an average cost per claim of AUD9,600 and an indirect cost estimated at five times that amount [1]. Whilst injury rates are slowly improving this breakdown appears to continue in developed countries on a global level. The physical, social and financial costs continue to remain at an unacceptably high level.

Workplace Health & Safety Standards in developed countries consistently require employers to provide their employees and contractors with a safe place to work. In relation to manual tasks, this is typically achieved by modifying tasks and equipment in an effort

[*] Corresponding author. Tel.: +61 7 4954 8652; Fax: +61 7 4954 8654; E-mail: jenny.legge@jobfitsystem.com.

to match the task to the worker. Sometimes, due to technical or cost considerations, this approach becomes impractical and the shift then changes to matching the worker to the task.

There have been a number of strategies employed to determine or attempt to minimize a worker's future risk of injury including back X-rays, manual handling training, history of previous pain and medical screenings including strength and endurance and body composition testing but there is limited evidence of their success [2,11,15]. A more recent approach in employee assessment is the use of pre-employment or post-offer functional assessments with the majority centered on the format of Functional Capacity Evaluations.

Functional Capacity Evaluations (FCEs), also commonly known as Functional Capacity Assessment or Physical Capacity Assessment, are typically constructed of a series of tests looking at the participant's mobility, strength (dynamic and isometric), cardiovascular fitness, tolerance to various positions and movements, as well as material handling ability including lifting, carrying, pushing and pulling. They also often include reports on the level of effort that the participant applied to the assessment. On most occasions, the results of the assessments are then compared to physical work demands either for determining a worker's ability to return to work following an injury, making recommendations in pre-employment or post-offer situations, monitoring progress during rehabilitation or for medicolegal and disability assessments and reporting [7]. A Pre-Employment Functional Assessment (PEFA) is a series of tests that provide objective information about a worker's functional capabilities in relation to the job for which they are applying.

Despite the limited published research examining the reliability and validity of functional capacity assessments, they have become a widely-used tool in the field of industrial rehabilitation. Of those that have been published many have focused on only one or two aspects of the assessment, such as fitness, strength or material handling [8]. None were identified that focused on all aspects included in this study nor allocated an overall performance score for comparative purposes. The JobFit System PEFA is based in parts on the WorkHab Functional Capacity Evaluation. This is the first scientific study investigating the reliability of the WorkHab FCE testing methods. Findings from other studies investigating the reliability of functional testing procedures are discussed below.

1.1. Purpose of the study

The purpose of this study is to determine the reliability of the JobFit System pre-employment functional assessments (PEFA) as a whole, or in parts, as a precursor for a validity study investigating the relationship between PEFA results and workplace injury rates and severity. With increasing pressure from all stakeholders (legal and health practitioners, workers and employers) the demand for evidence-based practice is rising. This reliability study and a subsequent validity study aim to meet those demands.

1.2. Reliability

Based on the National Institute for Occupational Safety and Health (NIOSH) criteria for the development and selection of work-related assessments there are five key attributes of an assessment: safety, reliability, validity, practicality and utility [6]. The issue of reliability is the subject of this study. Reliability refers to the level of consistency or repeatability between the measurements recorded for a test on different occasions (test-retest, intra-rater), and between different assessors (inter-rater). Clinically, this typically refers to obtaining the same results rather than proportional and consistent change [5].

1.2.1. Sources of error in reliability

Errors in measurement, and thus a reduction in reliability, can come from four major sources:

Participant – fatigue and health, motivation and attitude, practice and memory, experience and knowledge
Testing – clarity of instruction and adherence to procedure
Scoring – suitability of scoring method, experience, competence, familiarity and accuracy of scorers
Instrumentation – calibration and setup of equipment, suitability of assessment tools [16].

The factors cited above as affecting the participant could also be applied to the assessor. These human sources of error, that is the participant and assessor, could also be influenced by environmental factors such as time of day, temperature and humidity, noise, visibility and other distractions.

1.2.2. Test-retest reliability

Test-retest reliability is an indicator of the stability of a test. That is, the ability to produce the same results on two different occasions on the assumption that the measure being scored does not change over time. The time between the two testing occasions varies and is a balance between the need for rest, the desire to reduce memory or avoid changes in the conditions, in the case of this study, changes in health and fitness of the participant. Sources of error in test-retest reliability could be from all four listed above, but in comparison to inter- and intra-tester reliability, it is assumed that participant and instrumentation errors would be expected to be higher.

1.2.3. Inter-tester reliability

Thomas and Nelson [16] also describe inter-tester reliability as "objectivity – the degree to which different testers can obtain the same scores on the same participants", or conversely is a measure of the variation between testers. Testing and scoring would be the main sources of error with this measure of reliability. To address these sources of error the majority of commercially available functional capacity testing tools have detailed procedures with which practitioners must become competent before they become 'certified' assessors.

1.2.4. Intra-tester reliability

Intra-tester reliability measures the consistency of scoring for an individual assessor on two different occasions. It is a form of test-retest reliability, however errors are influenced more by testing and scoring rather than participant and instrumentation sources.

Intra- and inter-rater reliability are considered to be particularly important when using subjective observations as is often the case when using work-related assessments. Reliability in work-related assessments is critically important so that any changes recorded in a worker's performance can be attributed to actual changes in their level of physical function and not simply an error in measurement [5]. Standardization of the procedures and scoring systems is the key to reducing the 'subjectivity' and improving the objectivity (reliability) of the assessments.

1.2.5. Methods of measuring reliability

The degree of reliability, or consistency between two sets of scores, is typically expressed as a correlation coefficient. As the degree of variance between two sets of the same variable are being compared, intraclass correlation is the appropriate method [16]. The intraclass correlation coefficient (ICC) is a number between 0 and 1. The closer to one, the higher the stability. However, the range of scores and sample size also need to be considered when interpreting the results.

Whilst there a number of different measures for interpreting each form of reliability, for simplicity and to facilitate interpretation of the results, a single respected measure, ICC, will be used. Where questions arise as to the potential suitability of this measure, percentage agreement in the raw data will also be examined and findings discussed. A review of the literature indicates that whilst there is no definitive source, it appears to be accepted, that an ICC score of < 0.75 is poor to moderate and > 0.75 is good. Portney and Watkins in [5] suggest that a score above or equal to 0.90 is required for clinical application to ensure valid interpretation of the findings. Gross and Battie [4] and Reneman et al. [12] go one step further, rating an ICC > 0.90 as excellent.

1.2.6. Reliability literature

Despite the wide use of FCEs, there is limited published literature on the inter-, intra- and test-retest reliability of functional capacity evaluations. Of that which is available, the results indicate good reliability. Test-retest and intra-rater reliability are the most widely published. Those using ICC as an indicator of reliability are reported.

Gross and Battie [4] examined the inter-rater reliability and test-retest reliability of three lifting (floor to waist, waist to crown and horizontal) and three carrying (front, right and left side) tasks. A group of five experienced occupational therapists assessed twenty-eight subjects with lower back pain who were currently participating in a rehabilitation program. They achieved good to excellent results with inconsistencies in subject's performance cited as the greatest source of variability. This source of error, as previously discussed, is expected when examining test-retest reliability. Gross and Battie's methods for reducing rater bias have been adopted in this study whereby a primary assessor and secondary assessor were assigned. The primary assessor interacted with the participant and was responsible for the safety of the client and progressively increasing the weights. The secondary assessor and the primary assessor on the test-retest trials, when watching the videotaped performance (Gross and Battie's were live) indicated at the conclusion of each set of repetitions whether they would increase the weight or stop the testing. When the testing was stopped, the weight achieved was revealed. The reason for stopping the test was also recorded independently.

A comparative study by Reneman et al. [14] examined a group of twenty-eight healthy subjects and looked at test-retest reliability of twenty-eight activities but only nine were scored using ICC scores. These were: lifting low, lifting high, carry short, carry long, carry right, carry left, pushing static, pulling static and shuttle walk test. The remaining nineteen activities were either incomplete or used kappa values. An ICC of > 0.75 was considered acceptable. With the exception of static pushing and the shuttle walk test, seven scored an ICC > 0.84. Out of the remaining measures, only the forward bend test in standing scored an acceptable level of reliability using ICC, despite the vast majority being rated 'acceptable' based on kappa values and percentage agreement (eight were reportedly suitable for ICC rating). As predicted, there was less variation in performance with the healthy subjects. Patient behaviour, testing protocols and evaluator variation were cited as the reasons for diminished reliability

2. Method

2.1. Subjects

A Queensland Coal Mine agreed to participate in the study. A total of 28 healthy workers participated in a generic PEFA. Twenty of the participants participated in a second trial between one week and three months later. Demographic data including age and their usual role were collected. Before testing, each participant was required to sign a written consent form outlining: (i) the components of the assessment (ii) the risks and expectations of submaximal physical testing and the precautions that would be taken (iii) the purpose of the assessment and the use and disclosure of the collected information (iv) the opportunity to discontinue testing at any time. The consent form was designed to meet relevant medico-legal and privacy law requirements. The study was approved by the Ethics Officer of the School of Human Movement Studies, University of Queensland. Participants were screened for exclusion factors prior to commencement of the assessment. Exclusion factors included current injury, significant injury or surgery in the last six months, elevated blood pressure (resting systolic $>$ 160 mmHg or resting diastolic $>$ 95 mmHg) or specific medical advice.

2.2. Experimental design

2.2.1. Assessment process

The Pre-employment Functional Assessments (PEFAs) were generic assessments representative of those used for coal miners in labor-intensive roles as identified with the JobFit System. The JobFit System is a software database program that contains the key physical requirements of jobs and the physical capabilities of workers in a same-value format for immediate and objective comparison. Each task has been analyzed by a physiotherapist and the following information recorded: task overview; frequency and duration; working posture requirements; material handling requirements; and any other relevant information such as environmental considerations. Working posture requirements are described as 'Never', 'Occasional', 'Frequent' or 'Continuous' as per the widely recognized US Department of Labor's Dictionary of Occupational Titles. This data is entered into the JobFit System. A Job Summary is then formulated by the JobFit System for a job based on the combined requirements of the tasks required for that job. Postural requirements for each task that were considered to be a high risk for work-related musculoskeletal disorders and the key requirements for the job were identified for inclusion in the PEFA. Material handling requirements were also identified.

Each PEFA contained the following components and was delivered in the same sequence:

1. musculoskeletal screen
2. balance test (single leg stance on stable and unstable ground)
3. aerobic fitness test (3-minute Step Test)
4. postural tolerances (sustained Reaching forward, Reaching overhead, Stooping, Squatting, Climbing)
5. material handling tasks (progressive Floor to bench, Bench to shoulder, Bench to Overhead and Bilateral Carry using a functional method)

The musculoskeletal screen was included to screen for any current injuries or physical limitations to the requirements of the remainder of the assessment only. It was not included as a predictor of performance as its use for this purpose has been refuted by several studies [9,10]. The musculoskeletal screen included general range of motion, manual muscle strength testing and postural screening by a physiotherapist.

The procedures for each task were fully explained to the participants prior to the commencement of each activity. The fitness test, postural tolerance tasks and

Table 1
Definition of PEFA scores

Score	Definition
One	Has demonstrated the functional capacity to perform the proposed position as described with no restrictions
Two	Has demonstrated the functional capacity to perform the proposed position as described with minimal restrictions (specified)
Three	Has demonstrated the functional capacity to perform the proposed position as described with moderate restrictions (specified)
Four	Has not demonstrated the functional capacity to meet the inherent requirements of the proposed position as described

material handling tasks closely follow those of the WorkHab Functional Capacity Evaluation as outlined in their procedure manual [3].

PEFA Score: A PEFA Score is the overall score for the worker's performance in comparison with the physical requirements of the job for which they are applying. A worker can be scored one, two, three or four. Table 1 defines each score. The overall PEFA score was determined with the use of the JobFit System by comparing the worker's capabilities against the requirement of a task. If all requirements were met, the indicators and scores were green and they obtained a PEFA score of one. If not, then they were red and their record was analyzed further. They scored a two if their material handling capacity was within 15% of the requirement and/or they had a single minor postural tolerance limitation. They scored a three if their material handling capacity was more than 15% of the requirement and/or they had more than one minor or one moderate postural tolerance limitation. They scored a four if a gross mismatch was present. Fitness and balance test results had no direct bearing on the PEFA score but were collected to determine their reliability prior to being used in a subsequent validity study.

2.2.2. Trial groups

All twenty-eight participants completed the first trial. Twenty completed a second trial. Selection for the second trial was based on participant availability amongst those of whom one week had lapsed since their initial assessment and who had volunteered to participate in the second trial. Each live assessment was videotaped and conducted by the primary assessor (A1). After a minimum period of one week had lapsed, the primary assessor also watched the videos and rescored the assessments.

Each first and second trial video was watched by the second assessor (A2) and scored allowing a minimum one week period between watching the first and second trial videos.

The trial groups are summarised as follows:

- Intra-rater comparisons
  - A1 Trial 1 Live vs A1 Trial 1 Video ($n = 28$)
  - A1 Trial 2 Live vs A1 Trial 2 Video ($n = 20$)
- Inter-rater comparisons
  - A1 Trial 1 Live vs A2 Trial 1 Video ($n = 28$)
  - A1 Trial 2 Live vs A2 Trial 2 Video ($n = 20$)
  - A1 Trial 1 Video vs A2 Trial 1 Video ($n = 28$)
  - A1 Trial 2 Video vs A2 Trial 2 Video ($n = 20$)
- Test-retest comparisons
  - A1 Trial 1 Live vs Trial 2 Live ($n = 20$)

2.2.3. Assessors

The primary assessor was a registered physiotherapist with six years experience in conducting functional capacity evaluations, five years as a registered WorkHab FCE provider and a JobFit System functional assessment trainer. The second assessor was a registered occupational therapist with one year experience in conducting functional capacity evaluations all as a registered WorkHab FCE provider who had participated in the JobFit System functional assessment training program.

2.3. Data analysis

Intraclass correlation coefficient (ICC) and percentage agreement were used to measure test-retest, intra- and inter-rater reliability. ICC scores greater than 0.75 were interpreted as good and scores greater than 0.90 were interpreted as excellent [4,5,12]. Where disagreements occurred, raw data was examined in an effort to offer explanations for the variations.

3. Results

3.1. Subjects

The group consisted of 28 males aged 19 to 55 years (Mean: 35.5 yrs). Half were currently employed in

Table 2
Intraclass Correlation Coefficients (ICC) and Confidence Intervals for Overall PEFA Scores

Comparison	ICC	Lower limit	Upper limit
Intra-rater reliability [A1 live vs. A1 video ($n = 48$)]	0.94	0.90	0.96
Inter-rater reliability [A1 video vs. A2 video ($n = 48$)]	0.83	0.74	0.89
Inter-rater reliability [A1 live vs. A2 video ($n = 48$)	0.84	0.75	0.90
Test-retest reliability [A1 trial 1 vs. A2 trial 2 ($n = 20$)]	0.78	0.57	0.89

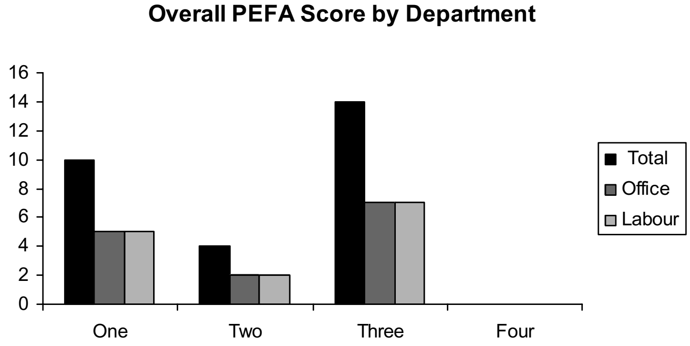


Fig. 1. Overall PEFA Score by Department.

an office/professional role (mean age: 36.1 yrs) and the other 50% were employed in a labor-intensive role (mean age: 34.9 yrs), the majority of which were underground coal miners. No subjects were excluded based on the musculoskeletal screen; however, one had temporary limitations identified in the lower limb due to pain from a recent tattoo.

3.2. PEFA score

The JobFit System PEFA score is determined by comparing a worker's capabilities to the job demands. The worker's material handling capacity is the primary factor. The second most influential factor is their postural tolerances. Fitness and balance test results do not have a significant effect on the overall score. The results for the various test components will thus be described in this order of influence rather than the order of data collection.

The PEFA scores for all participants by department are illustrated in Fig. 1. PEFA scores range from 1 to 4, with 1 being the better score. It is interesting to note that despite the huge variation in physical demands of their usual roles, on average, each group scored equally on the overall PEFA score. There were twice as many scoring 3 (moderate limitations) as there were scoring 1 or 2.

ICC scores indicate good to excellent reliability in determining the overall PEFA score (Table 2). One of the limitations of the ICC is that when only a small sample and small range of scores is used, a single change can have a dramatic result and can provide an inaccurate representation of the data. For this reason, actual values are discussed in the following paragraphs.

Test-retest: Twenty participants completed two trials. Sixteen (80%) of these showed consistency between trials. Three improved and one declined in performance. These are identified as participants 7, 10, 23 and 3 in Fig. 2.

Participant seven improved from a PEFA score of two to one. This was a direct result of increasing his overhead lifting capacity from 30.5 kg to 35 kg. It was noted, that the assessing therapist stopped the participant at 30.5 kg in the first trial, as it was determined that their safe lifting tolerance had been reached. The second assessor, when watching the video scored both trial one and trial two less at 23 kg and 30.5 kg respectively, again an improvement between trials albeit a more conservative score. Participant seven attributed his improvement to rugby training.

Participant ten also improved from a PEFA score of two to one also as a result of increasing their overhead lifting capacity from 30.5 kg to 35 kg. The result achieved in trial one was due to the participant stopping the test due to complaints of wrist discomfort. The second assessor did not agree with the improvement in trial two.

Participant twenty-three had the biggest improvement from three to one increasing his shoulder lift from 28 kg to 35 kg and his overhead lift from 23 kg to

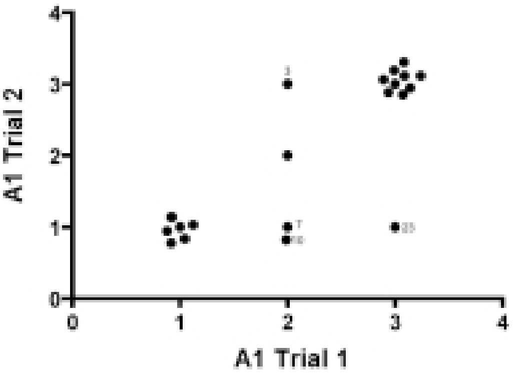


Fig. 2. Test-retest Reliability for Overall Score (Live Trials).

35 kg. No reason was documented for these improvements. Motivation, or fatigue in the first trial, is expected to be the main contributing factor as only two weeks had passed between trials thus making a training effect unlikely. Both assessors agreed on the original and revised scores.

Participant three who declined in his performance lowered his overhead lifting capacity from 30.5 kg to 28 kg. His shoulder lifting capacity also decreased from 33 kg to 30.5 kg but this would not have affected his overall score. Both assessors agreed on the change in results. There was no reason documented for his decline in performance between trials.

As was attributed by Gross and Battie [4], Reneman et al. [12,14] and Tuckwell et al. [17], participant variation appeared to be the main source of error.

When looking at the scatter plots in Figs 3 and 4, two clear trends appear:

1. the second assessor was consistently more conservative, and
2. video assessments were typically scored more conservatively than live assessments.

Inter-rater: Eleven of the forty-eight trials (23%) varied between assessors. The main differences between the assessors were years of experience and different disciplines. As both are looking for the same signs of safe maximal lifting and it is expected that each discipline would have equivalent observational skills, it is reasonable to assume that the main contributing factor would be confidence based perhaps on years of experience or personality differences. Reneman et al. [13] investigating the reliability of determining effort level of lifting and carrying in a functional capacity evaluation compared the inter-rater reliability of three physical therapists and two occupational therapists, four of which had only minimal experience. Reliability was expressed as a percentage and ranged from 87% to 96% which is a fair representation of the results achieved in this study. The variations between the different disciplines and the experience levels were not published and so could not be compared.

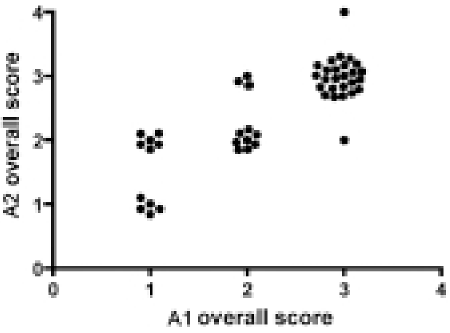


Fig. 3. Inter-rater Reliability for Overall PEFA Score (Video Trials).

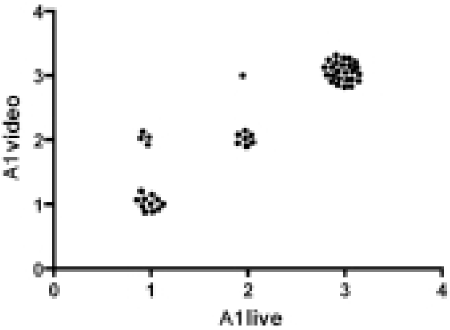
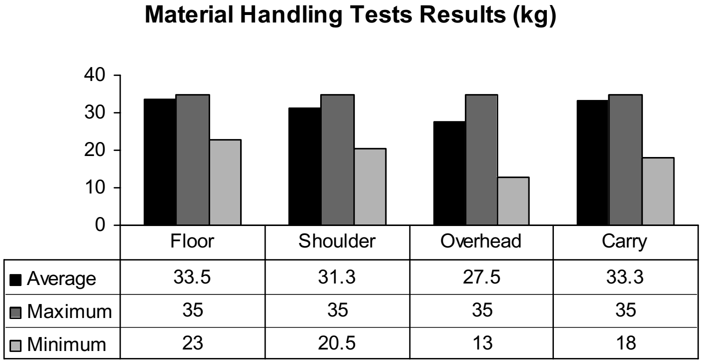


Fig. 4. Intra-rater Reliability for Overall PEFA Score (Trials 1 and 2).

Intra-rater: In the few cases that varied between live

Table 3
Intra-class Correlation Coefficient (ICC) scores and Confidence Intervals for Material Handling Tests

Test	Inter-rater [live vs. video ($n = 48$)]	Inter-rater [video vs. video ($n = 48$)]	Intra-rater [live vs. video ($n = 48$]
Floor to bench	0.96 (0.93–0.98)	0.98 (0.96–0.99)	0.98 (0.96–0.99)
Bench to shoulder	0.92 (0.87–0.95)	0.81 (0.70–0.88)	0.86 (0.78–0.91)
Bench to overhead	0.89 (0.83–0.93)	0.91 (0.85–0.94)	0.95 (0.93–0.97)
Bilateral carry	0.96 (0.94–0.98)	0.96 (0.94–0.98)	1.0



	Floor	Shoulder	Overhead	Carry
■ Average	33.5	31.3	27.5	33.3
▨ Maximum	35	35	35	35
□ Minimum	23	20.5	13	18

Fig. 5. Material Handling Tests Results.

and video scores, the video scores were typically rated lower. Three explanations are offered:

1. in the live scenario, the assessor can receive feedback from the participant when the decision to increase or stop is uncertain;
2. in the live scenario, the assessor can alter their observation point to obtain more information;
3. in the video situation, the assessor can pause for more time or rewind the tape if uncertain of the participant's performance.

Only five of the forty-eight trials (10%) varied for the first assessor.

3.3. *Material handling tests*

Four different material handling tests were conducted – floor to bench lift, bench to shoulder lift, bench to overhead lift and bilateral carry. Combining both trials, the average, high and low results for each are tabulated in Fig. 5.

Inter-rater ICC values ranged from 0.81 to 0.98 (good to excellent), and intra-rater ICC values ranged from 0.86 to 1 (good to excellent). Whilst the range of available scores with the material handling was larger than that of the postural tolerances and the confidence intervals overall much narrower, the use of the ICC for determining inter- and intra-rater reliability is still questionable (Table 3). The largest variation in these measures of reliability was with the bench to shoulder lifts. This could be due to the difficulty in observing the onset of compensatory movements and loss of postural control with this task in comparison to the others. Renemen et al. [14] also scored lower reliability on the 'high' lift compared to the 'low' lift but scored no difference in their earlier study [12]. An explanation for the lower score was not offered.

Test-retest ICC values ranged from 0.56 to 0.88 (poor to good) The sample size for the test-retest ($n = 19$ to 20) and the narrow range of results for the floor to bench and bench to shoulder lifts further weakened the value of determining the ICC for this group. These results have been included (Table 4) simply to illustrate this point. Discussion of the results in the following paragraphs will give a more accurate representation of the test-retest reliability and the implications that this would have on the participant's overall PEFA score.

3.3.1. *Floor to bench*

Test-retest: Only nineteen floor to bench trials were included, as one participant could not comfortably squat during the first trial due to discomfort from a recent tattoo. Only four scores (21%) varied between trials. The variation is illustrated in Fig. 6. Two improved and two declined in performance, both due to self-limiting behavior. That is, the worker stopped the test prematurely with complaints of lower back pain for one, and feeling 'heady with sinus' by the other. The

Table 4
Test-retest Intra-class Correlation Coefficient (ICC) Scores and Confidence Intervals for Material Handling Tests

Test	ICC	Lower limit	Upper limit
Floor to bench ($n = 19$)	0.56	0.22	0.78
Bench to shoulder ($n = 20$)	0.64	0.34	0.81
Bench to overhead ($n = 20$)	0.82	0.63	0.91
Bilateral carry ($n = 20$)	0.88	0.74	0.94

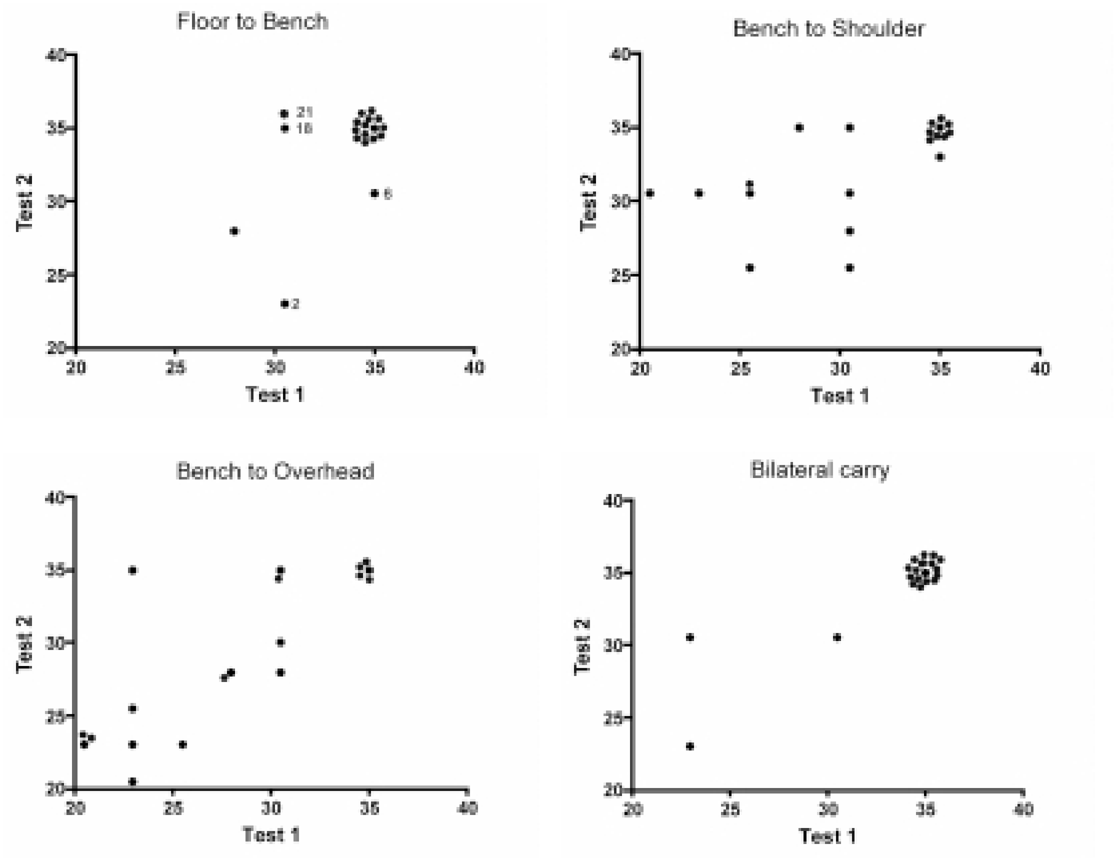


Fig. 6. Test-retest Reliability of Material Handling Tests.

worker with lower back pain declined in performance from 30 kg to 22 kg. This is a positive indicator of the validity of this assessment methodology. Both results would have lowered their overall score. In both cases, the intra-rater and inter-rater scores were 100% consistent. Conversely, the two participants that improved would have increased their score and similarly the intra-rater and inter-rater scores were in agreement.

Inter-rater: Of forty-seven trials, there were two variations in scores demonstrating excellent reliability.

Intra-rater: There was only one variation in scoring. This variation was agreed upon by both assessors watching the video.

3.3.2. Bench to shoulder

Test-retest: As indicated by the confidence intervals, the variation in bench to shoulder lifts was larger. Of the twenty trials, eight (40%) varied between trials. Only three declined in their performance. One of these was self-limiting, the other two were based on the assessors' decision. The second two only declined in performance by 2 kg. Two of the three would have achieved a lower overall score. The other five variations were improvements in performance, ranging from 5 to 7 kg. All of these would have achieved a higher overall score. It is suspected that motivation was a major contributing factor to this change.

Table 5
Intra-class Correlation Coefficient (ICC) Scores and Confidence Intervals for Postural Tolerances Tasks

Test	Inter-rater (live vs. video)	Inter-rater (video vs. video)	Intra-rater (live vs. video)
Reach Forward	0.87 (0.79–0.92)	0.93 (0.89–0.96)	0.93 (0.89–0.96)
Reach Overhead	0.86 (0.78–0.91)	0.75 (0.62–0.84)	0.60 (0.41–0.73)
Stoop	0.84 (0.75–0.90)	0.72 (0.57–0.82)	0.81 (0.70–0.88)
Squat	0.68 (0.53–0.80)	0.82 (0.72–0.89)	0.67 (0.51–0.78)
Climbing	1	1	1

Inter-rater: A quarter of the 48 trials recorded variation between the assessors, with the second assessor typically more conservative.

Intra-rater: 14.5% of the trials recorded an intra-rater variation with the video score typically more conservative than the live score.

3.3.3. Bench to overhead

Test-retest: Again, there was significant variation amongst the two trials for the bench to overhead lift. However, only three declined in performance with the results of only one affecting their overall score. As with seven of the ten variations, the change was only 2–2.5 kg which was one increment in the progressive weight protocol. It is worth noting that one participant improved from 23 kg to 35 kg which would have improved their score from a three to a one. The reason for this dramatic improvement is not known however, it was noted that they improved on all aspects of their test, excluding fitness.

Inter-rater: As predicted by the confidence intervals, the variation in scores between assessors was higher (16 out of 48, 33%) for the bench to overhead lift with the live assessor again giving higher scores

Intra-rater: The intra-rater variation (7 out of 48) was the same as for the bench to shoulder lift with no identifiable trend to lower scores on video or live.

3.3.4. Bilateral carry

Test-retest: Out of twenty participants, only one varied between trials. His improvement of 7 kg was directly as a result of self-limiting behaviour in the first trial. That is, he stopped the test prior to the assessor determining that his safe maximal lift had been reached. The improvement would have resulted in him achieving a higher PEFA score.

Inter-rater: Of the forty-eight trials, there were three occasions where the second assessor would have scored the participant one increment lower on the bilateral carry task.

Intra-rater: No variation recorded.

3.4. Postural and dynamic tolerances tests

As discussed previously, one of the limiting factors of using the ICC as a measure of reliability is that when there is only a small range in the values it loses some of its sensitivity. In these cases, such as the postural tolerances results below, reporting of individual scores and explanation of the variation from the raw data can provide more useful information. This limitation is magnified when a small sample size ($n = 20$ for test-retest) is also used. The ICC results for the inter-rater and intra-rater reliability for the postural tolerances are tabulated (Table 5) with more detailed explanations in the following paragraphs. No consistent trend between video vs. video and live vs. video was identified and so it can be assumed that the medium did not make a significant difference to the result in the postural tolerances tasks.

A review of the literature indicates considerable variation in reliability for postural tolerances tasks. Reneman et al. [14] reported high agreement and reliability for crouching, whereas Tuckwell et al. [17] reported lower readings similar to the trend in this study. Conversely, their stair climbing rated poorly compared to the results obtained in this study and that of Reneman's et al. [14].

3.4.1. Forward reach

Test-retest: Six of the twenty participants varied between trial one and trial two. Three improved from 'F' to 'X' and three decreased from 'X' to 'F'. Of those that decreased, two reported feeling unwell. The third's result was based solely on heart rates changes and was also scored inconsistently between the raters. These changes would not have changed their overall score.

Inter-rater: Of the forty-eight trials, there were only two variations. There was a 50/50 split between variation of live vs. video and video vs. video.

Intra-rater: There was only variation of the forty-eight trials which is indicative of excellent reliability.

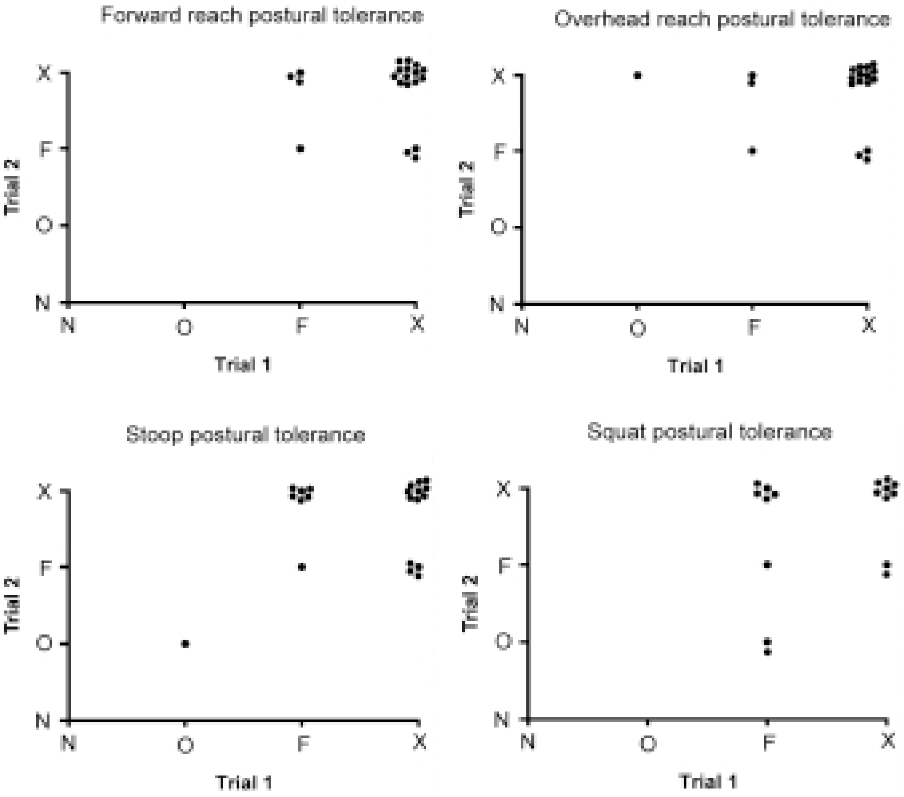


Fig. 7. Test-retest Reliability of Postural Tolerances Tests.

3.4.2. Overhead reach

Test-retest: Again, there were six variations between trials one and two. Three also improved, this time, two from 'F' to 'X' which would not have altered their overall score, but one from 'O' to 'X' which would have increased their score. Evaluation of the raw data demonstrated this participant did not complete the task in the first trial. This variation is therefore a positive indicator toward the validity of the data. The three participants whose score reduced from 'X' to 'F' all reported arm fatigue with corresponding changes in their heart rates. The workers reported no explanation for their change in performance. These scores would not have changed their overall rating but would indicate a referral for behaviour modification such as avoiding repetitive or sustained overhead reaching.

Inter-rater: Variation in this task was double that of forward reach (8% versus 4%) but still low.

Intra-rater: There were six variations amongst the forty-eight trials (12.5%). Although this is higher than the forward reach, this still indicates good reliability despite a moderate score in the ICC value (0.60).

3.4.3. Stoop

Test-retest: There was a higher rate of variation for the stooping task. Half of the results varied between trials but only four worsened. Three out of the four participants reported discomfort, two from football training the night before. Changes in heart rate coincided with three of the changes. Only one had disagreement between assessors. None of the changes would have affected the participant's overall score.

Inter-rater: Variation was the same as the overhead reach task (four of the forty-eight trials).

Intra-rater: Intra-rater variation was also the same as the overhead reach task, again scoring good. The ICC value in this case however was 0.81.

3.4.4. Squat

Test-retest: Eight of the nineteen participants (42%) varied between trials of squatting tolerance but with only three showing a decline in performance, one of which was due to self-limiting behaviour (i.e. stopped test prematurely). The other two decreased from an 'X' to an 'F'. Evaluation of the raw data shows that this was based on heart rate change alone and these scores did

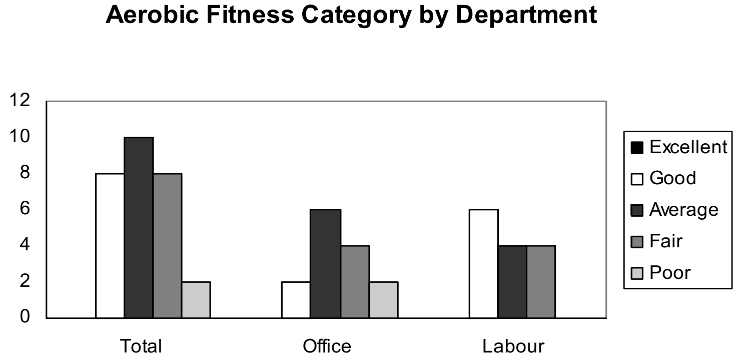


Fig. 8. Aerobic Fitness Category by Department.

not show intra- or inter-rater reliability. These results did not affect the participant's overall scores.

Inter-rater: Variation was highest in the squatting task. Six (12.5%) of the forty-eight trials varied.

Intra-rater: Intra-rater variability was also the highest at 14.5% (seven trials). These higher rates of variation could be contributed to less clear definition of compensatory behaviour. It could also indicate that heart rate changes during this task may not be as strong an indicator of discomfort or effort as large muscle groups are not being used and the task is performed lower to the ground thus decreasing the work of the heart.

3.4.5. Climbing

There was no variation in the climbing scores with the test-retest, inter-rater or intra-rater comparisons.

3.5. Fitness test

The results of the aerobic fitness test are illustrated in Fig. 8. Two participants did not complete the test within their 85% MHR and thus rated 'poor'. Nineteen fitness test results were recorded for both trials. Ten participants scored the same result in both trials (five fair, three average and two good). Three declined in their rating and four improved. It is worth noting, that whilst the two departments scored equally on the overall PEFA score, those employed in the labor-intensive roles, on average demonstrated higher levels of aerobic fitness by an increased number with a rating of 'good' (six versus two).

Test-retest scores for the fitness tests are illustrated below (Fig. 9). Due to the variation in results between trials one and two of the fitness test, recovery heart rates were also compared in an attempt to account for the variation. No clear and consistent explanation can be offered for these results. Factors influencing heart rates include, but are not limited to: emotional state, physical fitness, prior activity, caffeine, tobacco, prescription and non-prescription drugs and fatigue. Whilst this extreme variation in fitness test results does not have any direct implications on the overall PEFA score it may influence the conclusions that can be drawn from the subsequent validity study. Whilst there a number of published articles on the reliability of the fitness test results in determining aerobic capacity, a peer-reviewed paper investigating the test-retest reliability of the chosen fitness test could not be found.

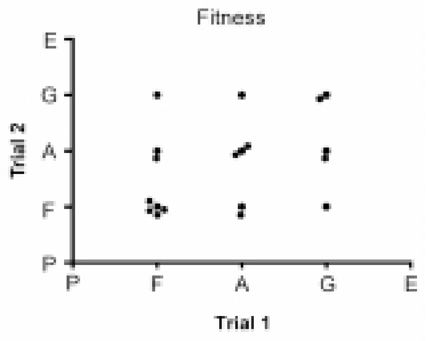


Fig. 9. Test-retest Reliability for the Fitness Test.

3.6. Balance test

Nineteen balance test results were recorded. Between the two trials, twelve participants consistently scored 'unlimited'. Two consistently scored 'limited'. Five scored a 'limited' result in trial one but improved to 'unlimited' in trial two. No reason for this improvement was documented nor reported by the participants. It is reasonable to assume that there is a positive practice and motivational component to the second trial results in these five participants.

Table 6
Reliability Ratings for PEFA Score and all Tests

Test	Test-retest	Inter-rater	Intra-rater
PEFA score	Good	Good	Excellent
Floor to bench lift	Moderate	Excellent	Excellent
Bench to shoulder lift	Moderate	Good	Good
Bench to overhead lift	Good	Good	Excellent
Bilateral carry	Good	Excellent	Excellent
Reaching Forward	Moderate	Good	Good
Reaching Overhead	Moderate	Good	Moderate
Stooping	Poor to moderate	Good	Good
Squatting	Poor to moderate	Moderate	Moderate
Climbing	Excellent	Excellent	Excellent
Fitness	Poor	NT	NT
Balance	Moderate	NT	NT

4. Discussion

Reliability encompasses test-retest, intra- and inter-rater reliability. Reliability of a measure needs to be determined prior to addressing the validity of a test. In consideration of the ICC values, confidence intervals and raw data, the reliability ratings for each test assessed in this study are tabulated below (Table 6).

As discussed previously, the ICC as a measure of reliability is not necessarily sensitive enough to account for the small ranges of values used in the components of this test. Therapists when interpreting these results for clinical use would be better informed by taking note of the actual values and reason for change between them rather than looking at the ICC alone. Due to variations in testing procedures and the use of different measures of reliability it is difficult to make comparisons between these results and other published papers, however there does seem to be some consistency between lower reliability scores for above shoulder lifts and tolerance to reaching forward and squatting.

This study was conducted at a working coal mine and therefore several limitations were not controlled. Variation in time between trials ranging from one week to two months existed. However, review of the data did not indicate an obvious effect from this variation. Participants were also exposed to variable levels of working hours, physical activity and mental stress immediately preceding their assessment. This is likely to have had an effect on their energy levels, concentration and heart rates. Differences in participant attitude is also likely to have had an effect. Participants were likely to be more relaxed on the second assessment which would have the potential to affect their heart rate and breathing patterns. Discussion of their performance, particularly manual handling tasks, with coworkers could have also resulted in an unintentional competitive environment which may have affected participant motivation on the second trial.

Despite the variation in some of the scores in this study, it was only a small number of cases where the changes would have affected the participant's overall score (six negatively, eight positively). The overall score is not meant to pass or fail potential job candidates but rather give the worker and the employer an indication of the level of risk of injury to that worker performing that role at that time. The individual test results are designed to offer both parties useful information on how the job can be modified or appropriate steps that the worker can take to minimize their risk of injury from manual handling injuries at work. The transference of the results of the PEFA into a workers' tolerance to a full day of work and avoidance of injury will be the basis for the subsequent validity study.

5. Conclusion

The overall PEFA score, climbing task and all four material handling tasks (floor to bench lift, bench to shoulder lift, bench to overhead lift and bilateral carry) demonstrated sufficient reliability for their inclusion in the subsequent validity study. The remaining tasks (excluding fitness) will be included but results will be interpreted with caution and will be weighted according to the reliability study findings. The fitness test results will not be used to draw conclusions in the validity study.

When interpreting these results, practitioners are reminded that 'excellence' in work-related assessments is achieved through a balancing act of the five key attributes – safety, reliability, validity, practicality and utility. It is generally accepted that a test is not deemed valid unless it is first considered reliable, yet as measures of reliability improve, measures of validity often decline. As a result, the practitioner must weigh up all the attributes when deciding which subtests to include and not base their decisions on the reliability or validity results alone.

References

[1] Australian Government: National Occupational Health & Safety Commission, *Compendium of Workers Compensation Statistics Australia, 2001–2002*. http://www.nohsc.gov.au /Statistics/publications/#compendium, Australian Government, 2003.

[2] S.J. Bigos and M.C. Battie, Preplacement Worker Testing and Selection Considerations, *Ergonomics* **30**(2) (1987), 249–251.

[3] S. Bradbury and D. Roberts, *WorkHab Australia Functional Capacity Profiling Procedure Manual*, Australia: WorkHab Australia, 1998.

[4] D.P. Gross and M.C. Battie, Reliability of Safe Maximum Lifting Determinations of a Functional Capacity Evaluation, *Physical Therapy* **82**(4) (2002), 364–371.

[5] E. Innes and L. Straker, Reliability of work-related assessments, *Work* **13** (1999), 107–124.

[6] E. Innes and L. Straker, Attributes of Excellence in Work-related Assessments, *Work* **20** (2003), 63–76.

[7] P.M. King, N. Tuckwell and T.E. Barrett , A Critical Review of Functional Capacity Evaluations, *Physical Therapy* **78**(8) (1998), 852–867.

[8] J. Legge, Pre-Employment Functional Assessments as an Effective Tool for Controlling Work-Related Musculoskeletal Disorders: A review, *Ergonomics Australia* **18**(2) (2004), 27–30.

[9] V. Mooney, K. Kenney, S. Leggett and B. Holmes, Relationship of Lumbar Strength in Shipyard Workers to Workplace Injury Claims, *Spine* **21**(17) (1996), 2001–2005.

[10] R.A. Mostardi, D.A. Noe, M.W. Kovacik and J.A. Porterfield, Isokinetic Lifting Strength and Occupational Injury: A Prospective Study, *Spine* **17**(2) (1992), 189–193.

[11] D.S. Reimer, B.D. Halbrook, P.H. Dreyfuss and C. Tibiletti, A Novel Approach to Preemployment Worker Fitness Evaluations in a Material-Handling Industry, *Spine* **19**(18) (1994), 2026–2032.

[12] M.F. Reneman, P.U. Dijkstra, M. Westmaas and L.N.H. Goeken, Test-Retest Reliability of Lifting and Carrying in a 2-day Functional Capacity Evaluation, *Journal of Occupational Rehabilitation* **12**(4) (2002), 269–275.

[13] M.F. Reneman, S.M.H.J. Jaegers, M. Westmass and L.N.H. Goeken, The reliability of determining effort level of lifting and carrying in a functional capacity evaluation, *Work* **18** (2002), 23–27.

[14] M.F. Reneman, S. Brouwer, A. Meinema, P.U. Dijkstra, J.H.B. Geertzen and J.W. Groothoff, Test-Retest Reliability of the Isernhagen Work Systems Functional Capacity Evaluation in Healthy Adults, *Journal of Occupational Rehabilitation* **14**(4) (2004), 295–305.

[15] S.H. Snook, Approaches to preplacement testing and selection of workers, *Ergonomics* **30**(2) (1987), 241–247.

[16] J.R. Thomas and J.K. Nelson, *Research Methods in Physical Activity,* 4th Edition, USA: Human Kinetics, 2001.

[17] N.L. Tuckwell, L. Straker and T.E. Barrett, Test-retest reliability on nine tasks of the Physical Work Performance Evaluation, *Work* **19** (2002), 243–253.