Queensland Mining Industry
2006 Health & Safety Conference 'Coping with Growth'

PRE-EMPLOYMENT AND PERIODIC FUNCTIONAL TESTING: A REVIEW OF THE EVIDENCE

Ms Jenny Legge BPhty MErg
JobFit Systems International, Mackay

Assoc Prof Robin Burgess-Limerick BHMS (Hons) PhD CPE
Burgess-Limerick & Associates, Brisbane

Keywords: Functional testing, Pre-employment, Workplace musculoskeletal injuries, Health surveillance, Injury Prevention

ABSTRACT

Functional testing has traditionally been the realm of therapists involved in end-stage rehabilitation or medicolegal claims for injured workers. The tide is turning. As the industry evolves and matures, functional testing is becoming more popular as a positive injury prevention, wellness and health surveillance tool to supplement other medical assessments. With the mining industry facing a competitive and ageing labour market, employers are recognising an increased need to better manage the health and wellness of their current workforce to improve both performance and retention.

Available research on functional testing is presented and discussed including the results of ACARP project C14045 investigating the reliability and validity of JobFit System pre-employment functional assessments. The use of functional testing for monitoring the changes in an ageing workforce and as an objective indicator of the need for task redesign to reduce the risk to our existing and future workers is also presented.

A matrix for evaluating the suitability of functional testing is introduced so that employers, insurers, workers and providers can weigh up their options and make informed decisions based on their individual priorities.

1. INTRODUCTION

It is estimated that sprains and strains are costing the Australian coal mining industry around $30 million per year in direct costs and about five times that amount in hidden or indirect costs. The social costs of workplace injuries also need to be considered. Whilst injury rates are slowly improving, around half continue to be from sprains and strains.

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 (Bigos & Battie 1987, Reimer *et al* 1994, Snook 1987, Mooney *et al* 1996, Mostardi *et al* 1992).

Functional capacity testing in the pre-employment or post-offer phase of recruitment is increasing in popularity as a risk management tool for controlling sprains and strains in the workplace and to assist in optimising performance and retention of existing workers. These assessments typically consist of a series of tests for mobility, strength, fitness, tolerance to different positions and movements, as well as material handling ability like lifting, carrying, pushing and pulling. Results are often compared to job demands to assist with decisions regarding job placement, task redesign and other risk management strategies such as physical conditioning programs.

Despite the limited published research examining the reliability and validity of functional capacity assessments they have become widely used (Legge 2004). With increasing pressure from all

stakeholders (legal and health practitioners, workers, insurers and employers) the demand for evidence-based practice is rising. This study which has been funded by ACARP aims to meet those demands.

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 (Innes & Straker 2003). Reliability relates to the level of consistency or repeatability between measurements and validity relates to its predictability and transferability to the workplace.

Reliability is commonly measured three ways. Test-retest reliability is an indicator of the stability of the test. Inter-tester reliability is often described as the objectivity of the test and intra-tester reliability refers to the consistency of the test. Practitioners using a reliable testing method can be confident that changes in performance on two different occasions can be attributed to change in the participant (eg. effort, motivation, conditioning) rather than as a result of variations in testing procedures and interpretation.

There is limited published research investigating the reliability of functional assessments, particularly in healthy workers which is the assumption in a pre-employment situation. Of those that were reviewed there appears to be some consistency between lower reliability scores for above shoulder lifts and tolerance to reaching forward and squatting, however there is a degree of variation between different functional capacity assessment methods (Gross & Battie 2002, Reneman *et al* 2004, Reneman *et al* 2002, Tuckwell 2002, Durand *et al* 2004).

Validity of pre-employment assessment tests can be measured by comparing test performance to injury rates, types of injuries, costs and duration of injuries, turnover rates and productivity. Of course, many factors influence these measures so whilst they are not the perfect indicators, they can provide reasonable feedback.

Gassoway and Flory (2000) conducted prework screens on 163 nursing assistants. A year later, they compared the turnover rates, injury costs and injury rates to a group of 144 nursing assistants hired in the previous year. The most dramatic result was reportedly the reduction in turnover rate from 60.4% to 41.7%. They also recorded reductions in the average injury costs from $ 377 to $ 320 and injury rates from 18.1% to 13.5%.

Nassau in 1999 in the 3rd stage of an injury prevention program, conducted 938 pre-work screens, of which 30 participants failed. When comparing the screened and non-screened group, Nassau identified that the average number of lost days in the screened group (0.83 days) was considerably lower than the unscreened group (3.83 days). Medical costs per 100 FTE was also lower at $311 compared to $1433 and the number of injuries similarly lower at 0.58 compared to 0.97.

2. METHODS

2.1 Reliability (secondary) Study

The purpose of this secondary study was 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.

A group of 28 healthy male coal mine employees participated in the study. Following the completion of an informed consent document and successful screening for exclusion factors, each worker was videotaped whilst participating in a generic PEFA representative of those used for coal miners in labour-intensive roles as identified with the JobFit System (Trial 1).

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

- musculoskeletal screen
- balance test
- fitness test
- postural tolerance activities (reaching forward, reaching overhead, squatting and stooping)

- material handling tasks (floor to bench lift, bench to shoulder lift, bench to overhead lift and bilateral carry).

At the completion of the assessment, the worker was given a PEFA score ranging between 1 (no limitations) and 4 (significant limitations).

Twenty of the participants completed the test a second time with a minimum of one week between trials (Trial 2).

Two assessors participated in the study – a physiotherapist and an occupational therapist. Each live assessment was scored and videotaped by the primary assessor (A1). After a minimum of one week, the assessor watched the videotapes and rescored the assessments. The second assessor (A2) watched the videotapes from both trials and scored the participants leaving a minimum of one week between watching the first and second trials.

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 (Gross & Battie 2002, Innes & Straker 1999, Reneman *et* al 2002). Where disagreements in scoring occurred, raw data was examined in an effort to offer explanations for the variations.

2.2 Primary (Validity) Study

The purpose of this primary study was to determine the validity of the JobFit System pre-employment functional assessments (PEFA) as a whole, or in parts by investigating the relationship between PEFA results and workplace injury rates and severity.

Since December 2002, new employees of a Queensland coal mine with both open cut and underground operations, have participated in a JobFit System pre-employment functional assessment. Their results have been recorded and retained for baseline data.

PEFA results have been compared to injury reports including the body part and nature of injury. Comparisons with severity and costs of injury will also be made.

3. RESULTS (Reliability Study)

3.1 Subjects

The study group consisted of 28 males aged 19 to 55 years (Mean: 35.5yrs). Half were currently employed in an office / professional role (Mean: 36.1yrs) and the other 50% were employed in labour-intensive roles (Mean: 34.9yrs), 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 was determined by comparing a worker's capabilities to the job demands. The worker's material handling capacity was the primary determining factor, followed by postural tolerances. Fitness and balance tests do not have a significant effect on the overall score. Scores range from 1 to 4 with 1 being the better score.

Ten participants scored 1 for their overall score, four scored 2 and the remainder (14) scored 3. Nobody scored 4. Despite the huge variation in the physical demands of their usual roles, on average, both groups scored equally on the overall score.

ICC scores indicated good to excellent reliability in determining the overall PEFA score. These are presented in Table 1. 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 provider an inaccurate representation of the data. For this reason, actual values were also considered when presenting the overall scores summarised in Table 3.

Table 1
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

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. Participants used a functional progressive lifting method until either their safe maximal lift was reached as determined by themselves or the assessor, or they reached the maximum requirement of 35kg. On average, participants lifted 33.5kg from floor to bench, 31.3kg from bench to shoulder, 27.5kg from bench to overhead and 33.3kg for the bilateral carry.

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). The largest variation in these measures of reliability was with the bench to shoulder lifts. This could be due to the perceived difficulty in observing the limiting factors with this task in comparison to others.

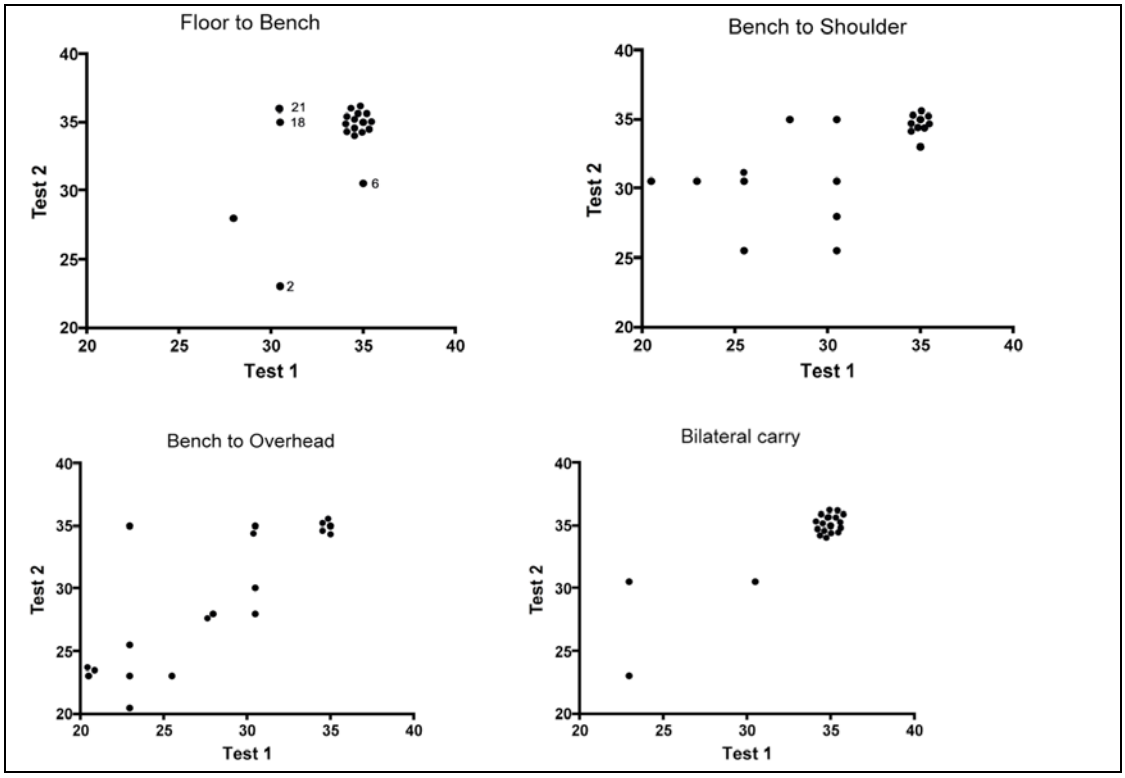

Figure 1. Test-retest Reliability of Material Handling Tests

Test-retest ICC values ranged from 0.56 to 0.88 (poor to good) with results illustrated in Figure 1. The sample size for the test-retest and the narrow range of results for the floor and shoulder lifts further weakened the value of determining ICC for this group. Investigation of the individual results gave a more accurate representation of the reliability and the implications on the overall score. This is discussed in detail in Legge & Burgess-Limerick (in press).

3.4 Postural and Dynamic Tolerances Tests

There were five postural and dynamic tolerance activities – reaching forward, reaching overhead, squatting, stooping and climbing. As discussed previously, the small range in values and the small sample size resulted in a dilution of the sensitivity of the ICC scores. The inter- and intra-rater reliability results are tabulated below (Table 2). Test-retest results are presented in Figure 2.

Table 2
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

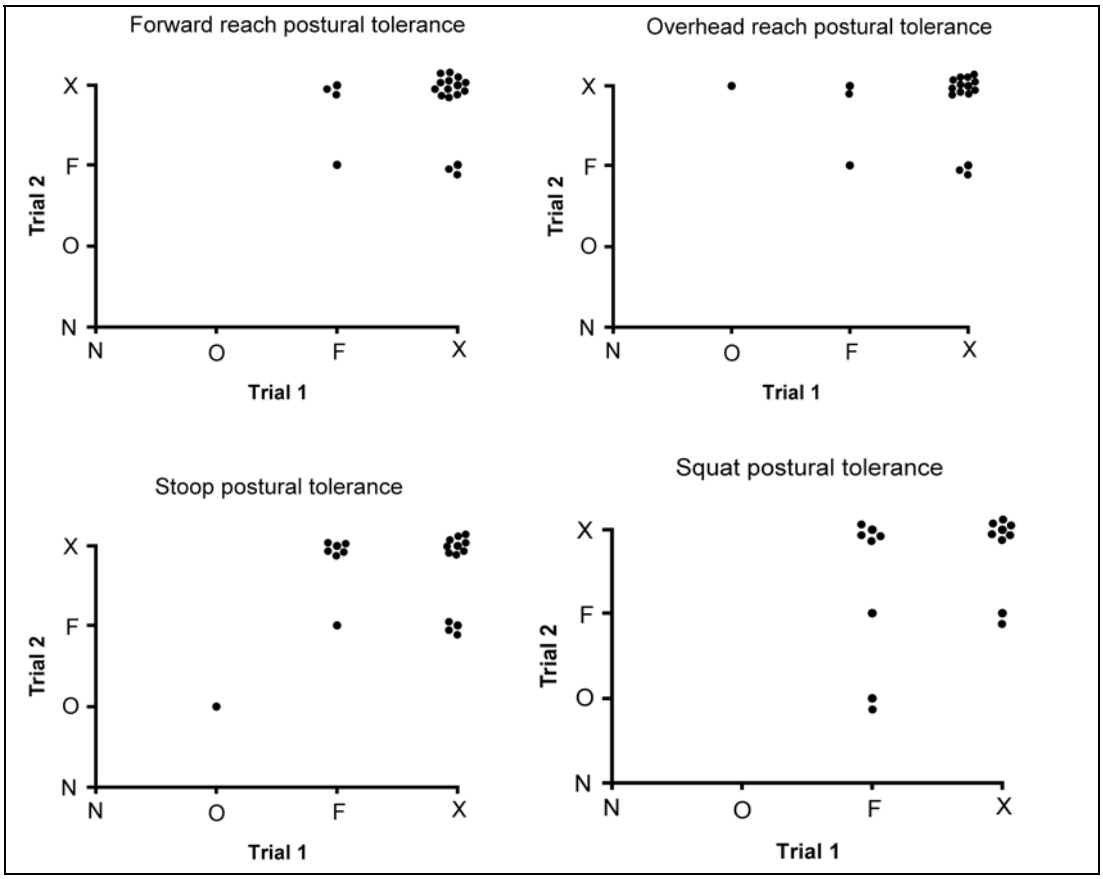

Figure 2. Test-retest Reliability of Postural Tolerances Tests

3.5 Fitness Test

The 3-minute Step Test chosen for this study, had the lowest test-retest reliability of all measures. Nineteen results were recorded and only ten participants scored the same in both trials. Three declined and four improved. No clear and consistent explanation can be offered for this variation and other peer-reviewed papers investigating the test-retest reliability of the step test were not found.

Factors that may have contributed to variations in heart rate and thus fitness rating include emotional state, physical fitness, prior activity, caffeine, tobacco, prescription and non-prescription drugs and fatigue. Whilst this extreme variation in fitness test results did not have an impact on the overall PEFA score, it may influence the conclusions that can be drawn from the subsequent validity study.

It is worth noting, that whilst the two departments (office and labour) scored equally on the PEFA score, on average those employed in labour-intensive roles scored higher fitness ratings.

3.6 Balance Test

Unilateral stance on stable and unstable ground with and without visual cues was used to test balance. Nineteen balance test results were recorded. Twelve scored ‘unlimited’ on both trials and

five improved. It is reasonable to assume that there is a positive practice and motivational component to the second trial results in these participants.

4. RESULTS (Validity Study)

Between December 2002 and January 2006, 403 job-specific pre-employment functional assessments have been conducted at the trial site. 304 of these people were hired and 240 records were suitable for analysis. Of this group, 187 scored 1, thirty-five scored 2 and eighteen scored 3.

Injury records coded as 'sprain / strain' were then compared to the PEFA results. Preliminary results will be presented in the oral presentation. To strengthen the statistical significance of this study, injury records will be collected for another 12months before being reported upon.

5. 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 3).

Table 3
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

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. One aspect that is clear though is that the reliability between different testing methods is low and therefore when comparing results within a group, the same standardized testing procedures need to be adopted for all.

The reliability 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 the reliability 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. They can also be used to monitor trends over time for health surveillance activities as the workforce ages and are exposed to musculoskeletal risk factors.

The transference of the results of the PEFA into a workers' tolerance to a full day of work and avoidance of injury was the basis for the subsequent validity study. The preliminary results of the study will be discussed in the oral presentation.

When deciding what type of pre-employment assessment to do and who to do it, decision-makers are advised to consider the five attributes of excellence for work-related assessments – safety, reliability, validity, practicality and utility. To assist with this process, a matrix that can be used to compare the three basic 'types' of assessments is presented (Figure 1).

	Clinical Physical	*Clinical Functional*	*Field Functional*
Safety			
Reliability			
Validity			
Practicality			
Utility			

Figure 3. Matrix for Evaluating Excellence in Work-Related Assessments

A 'clinical physical' assessment could be described as a traditional medical or musculoskeletal assessment performed by a medical practitioner or physiotherapist (eg. joint range of motion, strength testing). A 'clinical functional' would be the type of assessment used in this study – functional activities in a controlled clinical environment. A 'field functional' would be an assessment conducted at a workplace doing work simulation activities with actual equipment. There are pros and cons for each type.

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. By referring to the available literature, decision-makers can weigh up the capacities and limitations of different types of assessments depending upon their prioritisation of the five attributes. This will assist in choosing the 'right' assessment for their environment.

6. CONCLUSIONS

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.

Pre-employment functional assessments are not designed as 'culling' tools and they are not foolproof indicators of current or potential injury. They are however potentially a valuable risk management tool for controlling the number and costs of workplace sprains and strains. By providing baseline data about a worker's capacity which can be compared to their job demands, injury risk factors can be identified and therefore managed through task redesign, appropriate job placement and rotation and conditioning programs.

Pre-employment assessments are only one part of the risk management process. They need to be supplemented by ergonomic interventions, training and education, wellness and injury management programs, supportive management and a cooperative workforce.

7. ACKNOWLEDGEMENTS

This presentation and study is supported by ACARP funding (Project C14045) and supported by Xstrata Newlands Coal Mine.

8. REFERENCES

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

Durand, M., Loisel, P., Poitras, S., Mercier, R., Stock, S.R. & Lemaire, J., 2004, The Interrater Reliability of a Functional Capacity Evaluation: The Physical Work Performance Evaluation, *Journal of Occupational Rehabilitation,* ***14(2),*** 119-129.

Gassoway, J. & Flory, V., 2000, Prework screen: Is it helpful in reducing injuries and costs?. *Work* ***15,*** 101-106.

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

Innes, E. & Straker, L., 1999, Reliability of work-related assessments. *Work* ***13,*** 107-124.

Innes, E. & Straker, L., 2003, Attributes of Excellence in Work-related Assessments. *Work* ***20***, 63-76.

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

Legge, J. & Burgess-Limerick, R., in press, Reliability of the JobFit System Pre-Employment Functional Assessment Tool

Mooney, V., Kenney, K., Leggett, S. & Holmes, B., 1996, Relationship of Lumbar Strength in Shipyard Workers to Workplace Injury Claims. *Spine 21 (17),* pp 2001-2005

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

Nassau, D.W., 1999, The Effects of Prework Functional Screening on Lowering an Employer's Injury Rate, Medical Costs, and Lost Work Days. *Spine 24 (3),* pp 269-274

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

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

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

Tuckwell, N.L., Straker, L., Barrett, T.E., 2002, Test-retest reliability on nine tasks of the Physical Work Performance Evaluation. *Work* ***19,*** 243-253.