

 Open access • Journal Article • DOI:10.3109/09638288.2010.548896

## **Inter- and intra-rater reliability of the manual handling component of the WorkHab Functional Capacity Evaluation — [Source link](#)**

[Carole James](#), [Lynette Mackenzie](#), [Mike Capra](#)

**Institutions:** [RMIT University](#), [University of Sydney](#), [University of Queensland](#)

**Published on:** 23 Sep 2011 - [Disability and Rehabilitation](#) (Informa Healthcare)

**Topics:** [Intra-rater reliability](#) and [Reliability \(statistics\)](#)

Related papers:

- [Test-retest reliability of the manual handling component of the WorkHab functional capacity evaluation in healthy adults.](#)
- [An evaluation of the inter-rater and intra-rater reliability of OccuPro's functional capacity evaluation.](#)
- [Inter- and intra-rater reliability of the OCRA checklist method in video-recorded manual work tasks.](#)
- [Intra-rater and inter-rater reliability of six musculoskeletal preparticipatory screening tests.](#)
- [Reliability of the motor learning strategy rating instrument for children and youth with acquired brain injury.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/inter-and-intra-rater-reliability-of-the-manual-handling-yjnitpu5za>



NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

James CL, Mackenzie L, Capra M, "Inter- and intra-rater reliability of the manual handling component of the WorkHab Functional Capacity Evaluation. Originally published in Disability and Rehabilitation, 33, p. 1797-1804, (2011).

Available from: <http://dx.doi.org/10.3109/09638288.2010.548896>

Accessed from: <http://hdl.handle.net/1959.13/920740>

## **TITLE PAGE.**

### **Title:**

Inter and intra-rater reliability of the manual handling component of the WorkHab functional capacity evaluation.

### **Authors:**

Carole James<sup>1</sup>, Dr Lynette Mackenzie<sup>2</sup> and Professor Mike Capra<sup>3</sup>

### **Running Head:**

Inter and Intra-rater reliability of the WorkHab FCE.

### **Correspondence:**

Carole James,

School of Health Sciences, University of Newcastle, University Drive,  
Callaghan, NSW 2308, Australia.

Tel: +61 2 49 216632

Fax: +61 2 49 216984

Email: [Carole.James@newcastle.edu.au](mailto:Carole.James@newcastle.edu.au)

---

<sup>1</sup> School of Health Sciences, University of Newcastle, NSW, Australia. Email: [Carole.James@newcastle.edu.au](mailto:Carole.James@newcastle.edu.au)

<sup>2</sup> Faculty of Health Sciences, University of Sydney, NSW, Australia. Email: [Lynette.Mackenzie@sydney.edu.au](mailto:Lynette.Mackenzie@sydney.edu.au)

<sup>3</sup> Faculty of Science, University of Queensland, Brisbane, QLD Australia. Email: [m.capra@uq.edu.au](mailto:m.capra@uq.edu.au)

## **Abstract**

The WorkHab Functional Capacity Evaluation (FCE) is widely used in Australian workplace injury management and occupational rehabilitation arenas however there is a lack of published literature regarding its reliability and validity.

Purpose: This study investigated the intra and inter-rater reliability of the manual handling component of this FCE.

Method: A DVD was produced containing footage of the manual handling components of the WorkHab conducted with four injured workers. Therapist raters (n=17) who were trained and accredited in use of the WorkHab FCE scored these components and 14 raters re-evaluated them after approximately 2 weeks. Ratings were compared using Intraclass Correlation Coefficients (ICC's), paired sample t-tests (intra-rater), chi-squared (inter-rater), and percentage agreement.

Results: Intra-rater agreement was high with ICC's for the manual handling components and manual handling score showing excellent reliability (0.94 – 0.98) and good reliability for identification of the safe maximal lift (ICC: 0.81). Overall inter-rater agreement ranged from good to excellent for the manual handling components and safe maximal lift determination (ICC >0.9). Agreement for safe maximal lift identification was good.

Conclusions: Ratings demonstrated substantial levels of intra-rater and inter-rater reliability for the lifting components of the WorkHab FCE.

Keywords: Functional Capacity Evaluation, Work, WorkHab Functional Capacity Evaluation, Reliability, Lifting.

## **Introduction**

Functional Capacity Evaluations (FCE) are widely used in the area of work injury management and occupational rehabilitation to define an individual's functional work abilities or limitations (1-4). Lifting capacity is one component commonly assessed during an FCE (5). There are many different FCE's commercially available and clinics have also developed their own non- standardised, work specific FCE's (6-15).

There are several approaches to assessment of capacity that can be used in the FCE (16). These include: a metabolic approach based on quantification of physiological measures (oxygen uptake, physiological stress etc); a biomechanical approach which is based upon the ability to perform work within safe musculoskeletal or neuromuscular limits; a psychophysical approach which assesses the ability to perform work within perceived limitations (16); and a combination of these. The latter acknowledges the complex inter-relationship between physical, cognitive, behavioural and contextual functioning.

The WorkHab Functional Capacity Evaluation is an Australian standardised assessment used in occupational rehabilitation to determine functional capacity. It incorporates physiological and kinematic performance measures including heart rate, pain, perceived exertion and the observation of biomechanics(17). There is however, limited evidence on the psychometric properties of this assessment tool (18-21). In the current occupational rehabilitation climate, evidence for the use of specific assessment tools is required with appropriate decisions being based on accurate and meaningful FCE's being essential for successful occupational rehabilitation.

Evidence of psychometric properties of assessment tools provides confidence to consumers and informs best practice (16) .

Reliability of an assessment is considered a pre-requisite for accurate measurement and is defined as the extent to which a measure is consistent, free from error and demonstrates reproducibility over time. There are different aspects to reliability; test re-rest, inter and intra rater reliability. Inter-rater reliability has been identified as being best assessed when raters are able to measure a response during a single trial, however the use of one or more raters in a 'real life' situation may not always be possible (22). In the case of an FCE, space, procedure and information used to determine results, such as observation, physiological measures and client comment may impact upon the ability of more than one rater to be involved. This technique has been used to assess reliability in some FCE's (16, 23). Video recording of subjects performing activities provides an opportunity for multiple observers to rate the same performance with the same information and has been used to assess inter-rater reliability for several FCE's (21, 24-26).

This study aimed to evaluate the intra-rater and inter-rater reliability of the manual handling (lifting) component of the WorkHab FCE to provide evidence for its use in practice.

## **Method**

### **Study design**

This was a cross sectional study design using therapist raters who were trained and accredited in use of the WorkHab FCE. Four injured workers who underwent an

FCE, agreed to have the manual handling component of the evaluation video recorded. This provided a range of manual handling (floor to bench lifts, bench to bench lifts and bench to shoulder lifts) from the different evaluations to be used in the production of a DVD. The manual handling component of each FCE was divided into lifting segments. Each lifting segment represented three lift repetitions at one weight, and a total of 35 lifting segments were included on the DVD. The lift segments were presented in random order on the DVD. Therapist raters were asked to score each of the thirty five lift segments using the WorkHab manual handling scoring system (17) and to identify if the lift was at the individual's maximum ability.

## **Subjects**

Injured workers, who were undertaking an FCE as part of their occupational rehabilitation program in relation to return to work, were recruited from one rehabilitation provider accredited with WorkCover (NSW). Health professionals who were accredited providers of the WorkHab FCE from a variety of workplaces were invited to participate.

## **Instruments**

### **WorkHab**

The manual handling component of the WorkHab FCE uses a modular box system, which allows boxes to be stacked at various heights. The subject is instructed to lift the load box (initially empty) from beginning (e.g. bench) to end height (e.g. shoulder) and return. This is repeated three times before additional weight is added to the load box, until the safe maximum lift is reached. The WorkHab FCE assessor reviews the heart rate readings as the assessment progresses (objective

physiological measurements). They observe the biomechanics of the lift and record reported pain and ratings of client perceived exertion (effort). The assessor records the weight lifted and calculates a manual handling score. Stance, posture, leverage, torque and pacing comprise the manual handling score, which is based on the principles of safe manual handling, with each of these components being rated on a scale of 0-4 with '0' being no adherence and '4' being the highest safety score. The sum of the score for each components is recorded as the manual handling score for each subject (17). A higher score indicates more appropriate manual handling techniques are being used.

## **Procedures**

Following ethics approval from the University Human Research Ethics Committee, health professionals working for one rehabilitation provider discussed the research with injured workers, who were being referred for an FCE as part of their occupational rehabilitation, and provided the information sheet and consent form. Prior to commencing the FCE, each subject gave informed consent and signed a consent form for the manual handling component of the FCE to be video-taped. The FCE assessors who worked for the rehabilitation provider also gave informed consent and signed a consent form agreeing to the manual handling component of the FCE they were conducting to be video- taped.

Accredited providers of the WorkHab FCE were sent an email informing them of the research and asking for expressions of interest to be involved in rating the DVD. Those providers that responded to the researcher were sent an information sheet and consent form. Providers that consented to be part of the study were sent the first



DVD for rating. Following return of the first DVD and the rating score sheet, and after an approximate 2 week time-lag, a second DVD was sent for the provider to re-evaluate. The same manual handling lift segments were on the second DVD, however these were randomised in a different sequence to those on the first DVD.

### **Data analysis**

All data was entered into SPSS (version 18.0) for analysis. Descriptive analysis, Intraclass Correlation Coefficients (ICC's), 95% confidence intervals, paired sample t-tests (intra-rater), chi-square (inter-rater) and percentage agreement were calculated where appropriate. Intra-rater reliability, the level of agreement when the same therapist viewed the same clip on two different occasions, was calculated using the ICC – model 3. This is a mixed model where the rater is considered the fixed effect and the subjects are considered the random effect and is appropriate for measuring intra-rater reliability as the measurements of a single rater cannot be generalised to other raters. The ICC used to determine inter-rater reliability was Model 2 where both raters and subjects are considered random effects (22). An ICC of 0.90 or more was considered a measure of excellent reliability, an ICC of 0.75 – 0.90 was considered good and an ICC of less than 0.75 was considered moderate to poor [20, 21].

Analysis was conducted on data for each of the manual handling components, stance, posture, leverage, torque and pace, (using a mean of individual raters' scores for each component) for each individual worker.

Sample size was calculated using the Confidence Intervals for Proportions - the Bayesian Method using Beta Distribution, using the lifts from four injured workers based on the number of lift segments included in each FCE.

## **Results**

### **Participants.**

Four injured workers who were undergoing a WorkHab FCE as part of their occupational rehabilitation program consented to have the manual handling component of the FCE video-taped. Seventeen accredited providers of the WorkHab FCE completed the inter-rater reliability component of the study and 14 completed the follow up evaluation for the intra-rater reliability.

Eighty eight percent of the study sample were occupational therapists (n= 15) and 12% physiotherapists (n=2). The mean number of years of professional experience was 11.2 years (SD: 7.3; Range: 2-27 years); the mean number of years of FCE experience was 8 years (SD: 6.1; Range: 1-20 years) and the mean number of years conducting WorkHab FCE's was 4.7 years (SD: 2.7; Range: 1-9 years).

### **Intra-rater agreement**

The overall intra-rater agreement (n=14) for the manual handling score and for each of the manual handling components (stance, posture, leverage, torque and pace) showed excellent reliability (ICC: 0.91 – 0.97), (see table 1).

*Insert table 1 here*

Fifty two percent of all ratings were identical and 87% were within one score of each other on the 0-4 point scoring system.

Intra-rater percentage agreement for determining whether a lift was a safe maximal lift was moderate with 70% of all ratings being identical. The ICC for safe maximal lift showed a good level of reliability (ICC: 0.81).

Results calculated for each of the four injured workers (subjects) provides further information of the intra-rater agreement for the manual handling score with results showing good agreement (ICC:0.79 – 0.88), as seen in table 2.

*Insert table 2 here*

Results for the individual manual handling components ranged from poor to excellent (table 3). There were significant differences ( $p < 0.05$ ) in the posture and leverage ratings (means) for subject 2 and in the torque rating for subject 3. Both subject 2 and 3 also had significant differences in the manual handling score. Four of the raters (23%) scored the manual handling components for subject 2 significantly higher ( $p < 0.05$ ) at the second rating and 2 of the raters did so for subject 3.

*Insert table 3 here*

### **Inter-rater agreement**

Overall inter-rater agreement ( $n=17$ ) for the manual handling score and for each of the manual handling components ranged from good to excellent with ICC's ranging from 0.77 to 0.91 (See table 4).

*Insert table 4 here*

Inter-rater percentage agreement of scoring for the safe maximal lift was identical in 68% of cases, and the ICC for determining a safe maximal lift was 0.9, indicating an excellent level of reliability.

The inter-rater agreement for each subject ranged from good to excellent (table 5). No significant difference between raters was found using chi square analysis for each of the manual handling components.

*Insert table 5 here*

## **Discussion**

Reliability of assessment tools has been identified as important for assessing the abilities of injured individuals (19). This study investigated the intra and inter-rater reliability of the manual handling components of the WorkHab FCE. The intraclass correlation coefficient measures the agreement between pairs of observations, and is accepted as a measure of reliability in relation to the discriminative capacity of a test (22).

### **Intra-rater reliability findings**

The intra-rater agreement in this study, as shown with ICC values and percentage agreement, indicates that health professionals can make consistent judgements on the manual handling scoring system and when determining safe maximal lifts during a WorkHab FCE. When looking at the results for individual injured workers (subjects) the ICC values for the manual handling score (calculated from the sum of the components, stance, posture, leverage, torque and pace) was good. The different manual handling components; stance, leverage and pace also had good to excellent ICC values. Posture and torque, however, when analysed for each subject, had poor to good ICC values suggesting this was less reliably scored.

Posture is scored according to the maintenance of the normal lordosis throughout the assessment , with non maintenance resulting in lower scores; torque is scored according to the amount of rotation relative to the pelvis, particularly looking at spinal

twisting in the low back (17). Raters in this study saw video-footage of lift segments from two angles (rear and side) however did not have the three dimensional vision that would be present in the clinical setting and this may have impacted upon the scoring of these components. Another reason for this difference could be due to the operational definitions for posture and torque. This may not be as detailed or clear to assessors as the operational definitions of other components of the manual handling scoring system. Use of operational definitions has been shown to improve reliability between and within raters (16). Another possibility is that posture or torque is more complex and therefore more difficult to assess. It should be noted, however that difference in the posture scoring component did not reflect negatively upon the reliability of the overall manual handling score, nor did the difference in the torque score. The manual handling score is the aspect used most in clinical practice, rather than the individual components.

The intra-rater agreement for individual injured workers (subjects) for each of the manual handling components showed that injured worker (subject) 2 was rated significantly higher at the re-assessment for stance, posture and leverage suggesting raters perceived a better manual handling technique was being employed on the second viewing of the DVD. These results were influenced by significant increases in ratings by 4 individual raters. Injured worker (subject) 3 however was rated significantly lower at the re-assessment, specifically for torque suggesting raters perceived the lifting technique was less good at the second viewing of the DVD. Experienced providers of the WorkHab FCE were rating one testing occasion for each subject at two different times. When considered across the 14 ratings, the overall manual handling score for subjects 2 and 3 however, had good ICC values.

These findings demonstrate that the WorkHab FCE is a reliable measure when the same person acts as the assessor. This supports the findings of earlier research which found the test retest reliability of the WorkHab FCE was substantial (27). Similar types of lifting tasks evaluated in other FCE's such as Isernhagen, Job Fit, Physical work performance evaluation and Ergo-kit have reported substantial or acceptable levels of intra-rater reliability (9, 21, 26, 28), however direct comparison is not possible due to differences in testing procedures and measures of reliability used.

### **Evaluation of safe maximal lifts**

Safe maximal lifts are commonly determined during FCE's to enable guidelines to be developed for individuals for return to work safely. Raters in our study identified if each lift segment was a safe maximal lift or not. The intra-rater agreement for determining safe maximal lift was good (ICC: 0.81), with percentage agreement being 70% (moderate). The inter-rater agreement for determining safe maximal lift was similar with an ICC of 0.90 (excellent) and percentage agreement of 68% which is moderate. Isernhagen (24) used video analysis of lift segments to assess reliability of light, moderate and heavy lifts and found higher levels of intra and inter-rater reliability when analysing just light and heavy lifts, using a kappa statistic. In this study of the WorkHab FCE, raters were blinded to the sequence of the lift, and therefore were unable to consider the effects on lifting technique as weight increased. The WorkHab uses physiological and kinematic performance measures, including observation of biomechanics, for assessors to determine safe maximal lifts, and viewing the lifts out of sequence may have removed biomechanical cues typically used in the clinical setting. This may have impacted upon these results.

Further investigation of rater determination of safe maximal lift during the WorkHab FCE is recommended.

### **Inter-rater reliability findings**

The overall inter-rater reliability for the manual handling score and for each of its components (stance, posture, leverage, torque and pace) were good to excellent (ICC: 0.77 – 0.91). Confidence intervals for the manual handling score, stance, leverage and safe maximal lift were narrow suggesting high precision of these results (22). Although the ICC values for posture and torque were good, the confidence intervals were larger suggesting more variation in these results between raters. These findings are in line with other studies investigating inter-rater reliability of FCE's (9, 21, 23, 29, 30), although due to differences in testing protocols and reliability measures used, direct comparisons between these results and other published research is not possible.

The inter-rater reliability for each individual injured worker showed a range of ICC values indicating good to excellent reliability for stance and pace (0.76 to 0.92) and moderate to good reliability for posture, leverage, and torque (0.70 to 0.87). As with intra-rater reliability, posture and torque was less reliably scored and further research to investigate the clinical reasoning used in determining scores is recommended.

### **Relationship of findings to clinical practice**

In this study, the manual handling component of the FCE for the injured worker was divided into lift segments, with each lift segment representing three lift repetitions at the one weight. In clinical practice, the score for the manual handling components is determined after observing the whole lift – from minimum to safe maximum weight (17). The lift segments in our DVD were also randomised therefore raters were seeing each lift segment in isolation rather than as part of a whole lift. Randomising

the lift segments enabled an objective evaluation of each of the manual handling components using just the information provided, without reference to what had occurred before or after the lift for each subject. This added rigour to the study, although the process was not clinically realistic. In clinical practice other information is available to the assessor such as verbal cues, facial expressions, and knowledge about the subject in regards to type of injury, type of job or reason for completing the FCE. Clinical reasoning skills used by therapists (raters) include gathering information about different aspects of the situation, perceiving and interpreting the cues and then on the basis of this information and of relevant (therapist) knowledge making a judgement or assessment (31). In clinical practice, raters will therefore have more information to guide the clinical reasoning used in the determination of the manual handling component score. The use of pragmatic clinical reasoning by therapists conducting FCE's involves evaluating if an action is feasible in a given situation taking account of the context, resources, therapist knowledge, skills and interests and the wider organisational, socio-cultural and political considerations (32). Investigation of the nature and processes of clinical reasoning when conducting an FCE is an area that deserves further exploration, particularly investigating the aspects of clinical reasoning used to determine ratings for the different components of manual handling within the WorkHab FCE. Studies that have used real time to evaluate rater agreement (9, 23, 29, 30) (where this clinical information was available) and studies that used video recordings (where this information was not available) to evaluate rater agreement (21, 24, 26) have however, both found substantial or acceptable levels of inter-rater agreement.



This study had 17 raters in the inter rater reliability component and 14 raters in the intra rater reliability component, which is more than some previous studies of inter-rater reliability in FCE's (21, 26, 29). As more raters are included in a study, the chance of variation in ratings increases. Therefore, these findings indicate that the WorkHab FCE is a reliable measure to assess manual handling because different therapist-raters provide consistent ratings when assessing injured workers. Health professionals need to evaluate all the attributes of FCE's which includes safety, reliability, validity, practicality and utility, to ensure high quality standards of practice. Further studies investigating the validity of the WorkHab FCE are recommended.

### **Conclusion.**

This research investigated the intra-rater and inter-rater reliability of the manual handling scoring system, and safe maximal lift determination of the WorkHab FCE and included floor to bench, bench to bench and bench to shoulder lifts. The results found substantial levels of intra-rater agreement as shown with both ICC and percentage agreement results. The Inter-rater reliability results also showed substantial levels of reliability. In previous research it has been suggested that therapists are basing decisions on historical information and practices rather than evidence based practice and current research. The results of this research contribute to the growing evidence of FCE's in practice and will provide therapists with information to guide practice.

### **Acknowledgement.**

The authors would like to thank the injured workers who consented to have the manual handling component of the WorkHab FCE they were completing to be video-

taped. Thanks go also to the WorkHab accredited providers who volunteered to participate in this study. This study was made possible with the support of a University of Newcastle grant (No GO187308).

Declaration of interest: The authors report no conflict of interest. The authors are responsible for the content and writing of this paper. This study was made possible with the support of a University of Newcastle grant (No GO187308).

## References

1. Jundt J, King P. Work rehabilitation programs: a 1997 survey. *Work*. 1999;12:139-44.
2. King PM, Tuckwell NL, Barrett TE. A critical review of functional capacity evaluations. *Physical Therapy*. 1998;78(8):852.
3. Gibson L, Strong J. A review of functional capacity evaluation practice. *work*. 1997;9:3-11.
4. Lee G, Chan C, Hui-Chan C. Work profile and functional capacity of formwork carpenters at construction sites. *Disability and Rehabilitation*. 2001;23(1):9-14.
5. Jones T, Kumar S. Functional capacity evaluation of manual materials handlers: a review. *Disability and Rehabilitation*. 2003;25(4-5):179-91.
6. James C, Mackenzie L. Health Professional's Perceptions and Practices in Relation to Functional Capacity Evaluations: Results of a Quantitative Survey. *Journal of Occupational Rehabilitation*. 2009;19(2):203-11.
7. Durand M-J, Brassard B, Hong Q, Lemaire J, Loisel P. Responsiveness of the Physical Work Performance Evaluation, a Functional Capacity Evaluation, in Patients with Low Back Pain. *Journal of Occupational Rehabilitation*. 2008;18(1):58-67.
8. Boadella J, Sluiter j, Frings-Dresen MHW. Reliability of Upper Extremity Tests Measured by the Ergos Work Simulator: A Pilot Study. *Journal of Occupational Rehabilitation*. 2003 December;13(4):219-32.
9. Durand M-J, Loisel P, Poitras S, Mercier R, Stock S, Lemaire J. The interrater reliability of a Functional Capacity Evaluation: The Physical Work Performance Evaluation. *Journal of Occupational Rehabilitation*. 2004 June;14(2):119-29.
10. Jay M, Lamb J, Watson R, Young I, Fearon F, Alday J, et al. Sensitivity and specificity of the indicators of sincere effort of the EPIC lift capacity test on a previously injured population. *Spine*. 2000 Jun 1;25(11):1405-12.
11. Matheson L, Rogers L, Kaskutas V, Dakos M. Reliability and reactivity of three new functional assessment measures. *Work*. 2002;18:41-50.
12. Reneman M, Brouwer S, Meinema A, Dijkstra P, Geertzen J, Groothoff J. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in healthy adults. *Journal of Occupational Rehabilitation*. 2004 Dec2004///;14(4):295-305.
13. Ting W, Wessel J, Brintnell S, Maikala R, Bhambhani Y. Validity of the Baltimore Therapeutic Equipment Work Simulator in the measurement of lifting endurance in healthy men. *American Journal of Occupational Therapy*. 2001 Mar-Apr;55(2):184-90.

14. Goutteborge V, Wind H, Kuijer P, Frings-Dresen M. Reliability and Validity of Functional Capacity Evaluation methods: a systematic review with reference to Blankenship, Ergos work simulator, Ergo-Kit and Isernhagen work system. *International Archives of Occupational Environmental Health*. 2004 November 2004;77:527-37.
15. Cotton A, Schonstein E, Adams R. Use of Functional Capacity Evaluations by rehabilitation providers in NSW. *Work*. 2006;26:287-95.
16. Gardener L, McKenna K. Reliability of occupational therapists in determining safe, maximal lifting capacity. *Australian Occupational Therapy Journal*. 1999 Sep;46(3):110-19.
17. Bradbury S, Roberts D. *Workhab Functional Capacity Evaluation Procedural Manual* WorkHab Australia; 1998.
18. Innes E. Reliability and Validity of Functional Capacity Evaluations: An Update. *International Journal of Disability Management Research*. 2006;1(1):135-48.
19. Innes E, Straker L. Reliability of work-related assessments. *Work*. 1999;13(2):107-24.
20. Innes E, Straker L. Validity of work-related assessments. *Work*. 1999;13(2):125-52.
21. Legge J, Burgess-Limerick R. Reliability of the JobFit system Pre-employment Functional assessment tool. *Work*. 2007;28:299-312.
22. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, New Jersey: Pearson Prentice Hall; 2009.
23. Gross D, Battie M. Reliability of safe maximum lifting determinations of a functional capacity evaluation. *Physical Therapy*. 2002;82(4):364-72.
24. Isernhagen SJ, Hart DL, Matheson LM. Reliability of independent observer judgments of level of lift effort in a kinesio-physical functional capacity evaluation. *Work*. 1999 03;12(2):145-50.
25. Smith RL. Therapist's ability to identify safe maximal lifting in low back pain patients during functional capacity evaluation. *Journal of Orthopaedic and Sports Physical Therapy*. 1994;19(5):227-81.
26. Reneman MF, Jaegar SMHJ, Westmaas M, Goeken L. The reliability of determining effort level of lifting and carrying in a functional capacity evaluation. *Work*. 2002;18:23-7.
27. James C, Mackenzie L, Capra M. Test retest reliability of the manual handling component of the WorkHab functional capacity evaluation in healthy adults. *Disability & Rehabilitation*. 2010;32(22):1863-9.
28. Goutteborge V, Wind H, Kuijer PP, Sluiter JK, Frings-Dresen MH. Intra- and Interrater Reliability of the Ergo-Kit Functional Capacity Evaluation Method in Adults Without Musculoskeletal Complaints. *Archives of Physical Medicine and Rehabilitation*. 2005;86(12):2354-60.
29. Goutteborge V, Wind H, Kuijer PP, Sluiter JK, Frings-Dresen MH. Reliability and Agreement of 5 Ergo-Kit Functional Capacity Evaluation Lifting Tests in Subjects With Low Back Pain. *Archives of Physical Medicine and Rehabilitation*. 2006;87(10):1365-70.
30. Lechner DE, Jackson JR, Roth DL, Straaton KV. Reliability and validity of a newly developed test of physical work performance. *J Occup Med*. 1994 Sep;36(9):997-1004.
31. Hagedorn R. *Foundations for Practice in Occupational Therapy*. 3rd ed. London: Churchill Livingstone; 2001.
32. James C, MacKenzie L. The clinical utility of functional capacity evaluations: the opinion of health professionals working within occupational rehabilitation. *Work*. 2009;33(3):231-9.



Table 1: Intra-rater agreement: ICC for Manual Handling components (N=14).

<b>Manual Handling Component</b>	<b>ICC<sup>a</sup></b>	<b>95% CI<sup>b</sup> of ICC</b>	<b>Interpretation of ICC</b>
<b>Manual Handling Score</b>	0.97	0.94 : 0.99	Excellent
<b>Stance</b>	0.97	0.93 : 0.98	Excellent
<b>Posture</b>	0.91	0.82 : 0.96	Excellent
<b>Leverage</b>	0.97	0.94 : 0.99	Excellent
<b>Torque</b>	0.93	0.86 : 0.97	Excellent
<b>Pace</b>	0.96	0.91 : 0.98	Excellent
<b>Safe Maximal Lift</b>	0.81	0.58 : 0.94	Good

<sup>a</sup> ICC = Intraclass Correlation Coefficient (two way mixed).

<sup>b</sup> 95% CI = 95% Confidence interval.

Table 2: Intra-rater agreement: Manual Handling Score Results: means, difference, standard deviations 95% confidence intervals and ICC for each injured worker (subject). (N=14)

<b>Manual Handling Component</b>	<b>Mean 1<sup>a</sup> (SD 1)</b>	<b>Mean 2<sup>b</sup> (SD 2)</b>	<b>Mean Difference (SD difference)</b>	<b>95% CI<sup>c</sup> of difference</b>	<b>ICC<sup>d</sup> (95% CI of ICC)</b>	<b>Interpretation of ICC</b>
<b>Injured worker 1 (n=8 lifts)</b>	16.94 (1.61)	16.5 (2.01)	0.44 (1.16)	-0.24 to 1.11	0.88 (0.64 to 0.96)	Good
<b>Injured worker 2 (n=8 lifts)</b>	13.39 (2.14)	16.79 (1.87)	-3.4* (1.32)	-4.16 to -2.64	0.88 (0.62 to 0.96)	Good
<b>Injured worker 3 (n=9 lifts)</b>	18.03 (1.36)	16.63 (1.99)	-0.91* (1.40)	0.58 to 2.22	0.79 (0.34 to 0.93)	Good
<b>Injured worker 4 (n=10 lifts)</b>	16.47 (1.79)	16.51 (2.05)	-0.02 (1.31)	-0.78 to 0.73	0.87 (0.58 to 0.95)	Good

<sup>a</sup> Mean 1 = mean manual handling score from DVD 1

<sup>b</sup> Mean 2 = mean manual handling score from DVD 2

<sup>c</sup>95% CI = 95% Confidence interval.

<sup>d</sup>ICC = Intraclass correlation coefficient (two way mixed)

\*Significant (two tailed) at  $p < 0.05$  (paired sample t-test)

Table 3: Intra-rater agreement-Manual Handling Component Results: mean difference (time1 : time 2), standard deviation, 95% confidence intervals and ICC for each injured worker (subjects) (N=14).

	Stance		Posture		Leverage		Torque		Pace	
	Mean dif <sup>a</sup> (SD <sup>c</sup> ) (95% CI <sup>b</sup> of difference)	ICC <sup>d</sup> (95%CI)	Mean dif <sup>a</sup> (SD <sup>c</sup> ) (95% CI <sup>b</sup> of difference)	ICC <sup>d</sup> (95% CI)	Mean dif <sup>a</sup> (SD <sup>c</sup> ) (95% CI <sup>b</sup> of difference)	ICC <sup>d</sup> (95% CI)	Mean dif <sup>a</sup> (SD <sup>c</sup> ) (95% CI <sup>b</sup> of difference)	ICC <sup>d</sup> (95% CI)	Mean dif <sup>a</sup> (SD <sup>c</sup> ) (95% CI <sup>b</sup> of difference)	ICC <sup>d</sup> (95% CI)
<b>Injured worker 1 (n=8 lifts)</b>	0.12 (0.35) (-0.07 to 0.32)	0.91 (0.71 to 0.97)	0.12 (0.32) (-0.06 to 0.31)	0.85 (0.55 to 0.95)	0.08 (0.32) (-0.10 to 0.26)	0.87 (0.59 to 0.95)	-0.07 (0.31) (-0.25 to 0.10)	0.79 (0.35 to 0.93)	0.18 (0.27) (0.01 to 0.33)	0.84 (0.51 to 0.95)
<b>Injured worker 2 (n=8 lifts)</b>	-1.06 (0.51)* (-1.35 to -0.76)	0.81 (0.39 to 0.94)	-1.05 (0.52)* (-1.35 to -0.75)	0.3 (-1.19 to 0.77)	-0.81 (0.42)* (-1.05 to -0.56)	0.84 (0.49 to 0.94)	-0.23 (0.49) (-0.51 to 0.05)	0.70 (0.06 to 0.90)	0.24 (0.31) (-0.4 to -0.05)	0.84 (0.51 to 0.95)
<b>Injured worker 3 (n=9 lifts)</b>	-0.08 (0.41) (-0.32 to 0.15)	0.88 (0.64 to 0.96)	0.32 (0.45) (0.06 to 0.58)	0.52 (-0.48 to 0.84)	0.22 (0.37) (0.01 to 0.43)	0.85 (0.52 to 0.95)	0.61 (0.37)* (0.39 to 0.82)	0.46 (-0.67 to 0.83)	0.33 (0.28)* (0.16 to 0.5)	0.79 (0.36 to 0.93)
<b>Injured worker 4 (n=10 lifts)</b>	0.12 (0.44) (-0.12 to 0.38)	0.87 (0.76 to 0.94)	0.08 (0.45) (-0.17 to 0.35)	0.83 (0.68 to 0.93)	0.11 (0.27) (-0.05 to 0.26)	0.87 (0.75 to 0.94)	-0.16 (0.27) (-0.32 to -0.01)	0.82 (0.66 to 0.92)	-0.18 (0.33) (-0.38 to 0.1)	0.8 (0.62 to 0.92)

<sup>a</sup> Mean difference = mean difference of score in DVD 1 and DVD 2

<sup>b</sup>95% CI of difference – 95% Confidence interval of the mean difference

<sup>c</sup>SD = Standard deviation

<sup>d</sup>ICC = Intraclass correlation coefficient (two way mixed).

\*Significant (two tailed) at p<0.05 (paired sample t-test)

Table 4: Inter-rater agreement: ICC for Manual Handling components (N=17).

<b>Manual Handling Component</b>	<b>ICC<sup>a</sup></b>	<b>95% CI<sup>b</sup> of ICC</b>	<b>Interpretation of ICC</b>
<b>Manual Handling Score</b>	0.90	0.79 : 0.96	Excellent
<b>Stance</b>	0.91	0.81 : 0.96	Excellent
<b>Posture</b>	0.79	0.58 : 0.92	Good
<b>Leverage</b>	0.91	0.81 : 0.96	Good
<b>Torque</b>	0.77	0.53 : 0.91	Good
<b>Pace</b>	0.82	0.63 : 0.93	Good
<b>Safe Maximal Lift</b>	0.9	0.80 : 0.96	Excellent

<sup>a</sup> ICC = Intraclass Correlation Coefficient (two way random).

<sup>b</sup> 95% CI = 95% Confidence interval.



Table 5. Inter-rater agreement for manual handling components for each injured worker (N=17).

	Stance		Posture		Leverage		Torque		Pace	
	Mean <sup>a</sup> (SD <sup>c</sup> ) 95% CI <sup>b</sup> of mean	ICC <sup>d</sup> (95%CI)	Mean <sup>a</sup> (SD <sup>c</sup> ) 95% CI <sup>b</sup> of mean	ICC <sup>d</sup> (95% CI)	Mean <sup>a</sup> (SD <sup>c</sup> ) 95% CI <sup>b</sup> of mean	ICC <sup>d</sup> (95% CI)	Mean <sup>a</sup> (SD <sup>c</sup> ) 95% CI <sup>b</sup> of mean	ICC <sup>d</sup> (95% CI)	Mean <sup>a</sup> (SD <sup>c</sup> ) 95% CI <sup>b</sup> of mean	ICC <sup>d</sup> (95% CI)
Injured worker 1 (n=8 lifts)	3.21 0.62 2.88 :3.52	0.91 (0.82 to 0.96)	3.13 0.47 2.88 :3.37	0.84 (0.69 to 0.93)	3.32 0.39 3.11: 3.51	0.70 (0.42 to 0.87)	3.67 0.34 3.49 :3.85	0.84 (0.7 to 0.93)	3.75 0.27 3.61 :3.89	0.76 (0.53 to 0.89)
Injured worker 2 (n=10 lifts)	2.26 0.69 1.91 :2.62	0.92 (0.84 to 0.96)	2.30 0.46 2.06: 2.53	0.74 (0.49 to 0.89)	2.55 0.60 2.24 :2.86	0.85 (0.71 to 0.94)	3.08 0.58 2.77: 3.38	0.83 (0.68 to 0.93)	3.38 0.42 3.36 :3.40	0.81 (0.63 to 0.92)
Injured worker 3 (n=8 lifts)	3.22 0.68 2.87:3.58	0.92 (0.85 to 0.96)	3.56 0.37 3.36: 3.75	0.79 (0.59 to 0.91)	3.55 0.43 3.33 :3.77	0.83 (0.67 to 0.93)	3.87 0.22 3.76 :3.98	0.78 (0.56 to 0.91)	3.8 0.26 3.66 :3.93	0.85 (0.71 to 0.94)
Injured worker 4 (n=9 lifts)	3.21 0.50 2.95 :3.46	0.87 (0.75 to 0.95)	3.19 0.42 2.97: 3.41	0.83 (0.68 to 0.93)	3.37 0.45 3.14 :3.61	0.87 (0.75 to 0.95)	3.42 0.33 3.25 :3.60	0.82 (0.66 to 0.93)	3.28 0.38 3.08 :3.48	0.80 (0.62 to 0.92)

<sup>a</sup> Mean = Mean score for each of the manual handling components.

<sup>b</sup> Standard deviation = standard deviation of mean score

<sup>c</sup>95% CI = 95% Confidence interval of mean scores.

<sup>d</sup>ICC = Intraclass correlation coefficient (two way random).