

James Carole Lesley, Mackenzie Lynette Ann, Capra Mike, '**Test-retest reliability of the manual handling component of the WorkHab functional capacity evaluation in healthy adults**', Disability and Rehabilitation, 32 1863-1869 (2010)

TITLE PAGE.

Title:

Test - retest reliability of the manual handling component of the WorkHab functional capacity evaluation in healthy adults.

Authors:

Carole James¹, Dr Lynette Mackenzie² and Professor Mike Capra³

Running Head:

Test - retest 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

Abstract.

¹ School of Health Sciences, University of Newcastle, NSW, Australia. Email: Carole.James@newcastle.edu.au

² School of Health Sciences, University of Newcastle, NSW, Australia. Email: l.mackenzie@usyd.edu.au

³ Faculty of Health Sciences, University of Queensland, Brisbane, QLD Australia. Email: m.capra@uq.edu.au

Purpose: The WorkHab Functional Capacity Evaluation (FCE) is one of many FCEs currently available and is widely used in the Australian workplace injury management and occupational rehabilitation arena. This study investigated the test-retest reliability of manual handling tasks within the WorkHab Functional Capacity Evaluation (FCE) in healthy adults.

Method: A convenience sample of 25 healthy subjects, consisting of 19 women and 6 men with a mean age of 29 years (SD: 12.0) participated in this study. Two FCE sessions were held a week apart and subjects completed a floor to bench, bench to shoulder and bench to bench lift. Analysis of the outcomes of the FCE included; descriptive analysis; Intra Class Correlations (ICC); kappa; percentage agreement; and 95% limits of agreement where appropriate.

Results: The ICC's for the three lifts show an excellent reliability (0.90 – 0.92), and a moderate reliability for the manual handling score (0.74). Further analysis of the components of the manual handling score found the percentage agreement was high for all components ranging from 72-92%, however the kappa scores suggested poor to moderate reliability (range -0.06 to 0.52). Internal consistency of the manual handling score was good (Cronbach's Alpha = 0.92) indicating this is a reliable scale.

Conclusions: The ratings for the lifting components identified substantial levels of test-retest reliability for the lifting components of the WorkHab FCE in healthy adults.

Keywords: Functional Capacity Evaluation, Occupational Rehabilitation, WorkHab Functional Capacity Evaluation, Test-retest reliability.

INTRODUCTION.

Functional capacity evaluations (FCEs) are an integral part of work injury prevention and occupational rehabilitation. FCEs are designed to provide a comprehensive, performance based assessment, to define the functional abilities and limitations of an individual in the context of safe, productive work tasks and are used for work fitness determinations and to facilitate return to work [1-4]. Functional capacity evaluations are commonly used with individuals, who have suffered work related musculo-skeletal injuries, as part of the rehabilitation process, and can be used for a range of injury and disease types. FCE's have been broadly categorised into three groups; for those with and without a specific job; and a job or work capacity evaluation [5, 6]. There are a variety of FCEs commercially available [7-15]. The WorkHab functional capacity evaluation is a popular assessment in the Australian occupational rehabilitation arena [16-18]. It is used in any of the three categories of FCE mentioned above, however, there is limited published literature on it's psychometric properties [19, 20]. The usefulness of an assessment tool depends upon the extent that health professionals and those requesting (and paying for) services can rely upon data as being accurate and meaningful[21]. Without evidence of psychometric properties including reliability and validity the confidence in a tool can be questioned. Evidence based practice is encouraged when treating or working with clients, however in relation to the WorkHab FCE there is limited published evidence available to inform practice, despite it's wide spread use.[17-20, 22].

Portney and Watkins (2009) identify reliability of an assessment as the first prerequisite when considering measurement and a precursor to determining validity. Reliability is the extent to which a measure is consistent, free from error, and demonstrates the reproducibility or dependability of the assessment over time [21]. Test-retest reliability indicates stability of the assessment and determines the consistency of measures from one testing occasion to another, on the assumption that the behaviour being scored does not change over time[21]. Test-retest reliability can be influenced by: testing effects (practice); rater bias; and the test-retest interval or time between the two testings.

Healthy adult populations have been used by various researchers in determining measurement properties of FCE's [12, 13, 23] and are often chosen for convenience. Results of test-retest reliability studies may also be affected by changes in the medical status of injured workers' between testing sessions. Access to an injured worker population for research is not always easy [24] and in Australia this is difficult due to the litigious nature of the workers compensation system.

This study aimed to evaluate the test-retest reliability for the floor to bench, bench to shoulder and bench to bench lifts of the WorkHab FCE in healthy adults. The study also aimed to investigate the test-retest reliability of the manual handling score given as part of the WorkHab FCE and to investigate the internal consistency of this scoring system. This research is an important contribution to evidence based practice for therapists using the WorkHab FCE in practice.

METHOD:

Subjects.

A convenience sample of 25 healthy adult volunteers recruited from a University staff and student population participated in this study. The study sample consisted of 19 women and 6 men ranging in age from 19 years to 54 years, with a mean age of 29 years (SD :12.0). The mean weight was 66.4kg (SD: 10.4) and the mean height was 167.9cm (SD: 8.4).

WorkHab FCE.

The WorkHab FCE is based on objective physiological measures, observations of biomechanics, reported pain and ratings of client perceived exertion (effort). The procedures for the manual handling component of the WorkHab FCE uses a modular box system, which allows boxes to be stacked at various heights. Each lift is fully explained and demonstrated to the subject prior to commencement. Boxes are set at an appropriate height, and 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. The assessment uses a protocol of increasing load at each height until the safe maximum lifting limit is reached. Baseline heart rate is taken initially and then heart rate readings are taken after each three lift set. During the WorkHab FCE, if a manual handling technique is observed to be poor, the assessor should educate the subject before proceeding [25]. Termination of the assessment can occur as a result of any of the following: 1). the subject choosing to cease the activity; 2). if the subjects' heart rate, as determined by the heart rate

monitor worn throughout the assessment, exceeds predetermined levels of age predicted maximal heart rate [26]; or 3). the assessor terminates the activity if lifting becomes unsafe. The assessor records the results of the weights lifted in kilograms after each component. The assessor also calculates the Manual handling score, out of 20. This is calculated from the manual handling components of stance, posture, leverage, torque and pacing using the scale of 0-4 with '0' being no adherence and '4' being the highest score. The higher the score, the more appropriate the manual handling technique used during the lifting component of the FCE.

Procedures.

Ethics was obtained from the University Human Research Ethics Committee, following which, subjects were recruited using posters to advertise the study and via an email sent from the School of Health Sciences office at the University inviting participation. Interested persons contacted the researcher directly to discuss the study, receive an information statement and subsequently arrange an assessment time.

Prior to commencing the FCE, each subject gave informed consent and signed a consent form. Each subject completed a pre-assessment screening, including: completion of a questionnaire to determine medical status; a musculoskeletal evaluation and blood pressure check, to determine any medical risks and to screen for current injuries -this was not used as a predictor of performance. Subjects also completed a 3 minute step test (aerobic fitness test) to determine heart rate recovery times prior to undertaking the manual handling component of the assessment.

For this study the height of the lift was relative to the subject's waist (for the bench components) and shoulder (for the shoulder component). The manual handling component / safe maximal lifting limits were determined for a floor to bench lift, a bench to shoulder lift and a bench to bench lift. Two testing sessions were held a week apart with the time of day being kept constant where possible. The FCE was conducted by one Occupational Therapist who was trained and accredited as a WorkHab assessor, with ten years experience in conducting FCEs. The subjects were asked to perform to their maximum abilities. Subjects were not provided with information on the results of session 1 until after session 2 was completed. Following session 1, the assessor was not given access to these results again until the conclusion of session 2.

Data analysis.

All data was entered into SPSS (version 16.0) for analysis. Descriptive analysis, one way random Intraclass Correlation Coefficients (ICC's), 95% confidence intervals, limits of agreement [27], paired sample t-test, kappa (weighted for ordinal data) and percentage agreement were calculated where appropriate. A ratio between the limits of agreement and the mean score was also calculated using the following formula $(1.96 \times \text{standard deviation of mean difference}) / \text{mean session 1 and 2} \times 100\%$. Percentage agreement, as a measure of agreement, can be used to determine reliability and kappa is a chance corrected measure of agreement considering both the proportion of observed agreements and the proportion expected by chance [28]. The internal consistency of the components of the FCE manual handling scoring

system were calculated using Cronbach's alpha.

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]. A kappa score of more than 0.80 represents excellent agreement, above 0.60 represents substantial levels of agreement; from 0.40 – 0.60 represents moderate agreement and below 0.40 poor to fair agreement [21]. A Cronbach's alpha of between 0.70 and 0.90 was considered to indicate sufficient internal consistency and indicates the items within the scale are measuring the same construct and can be considered reliable [28, 29].

RESULTS.

The means, standard deviations, limits of agreement, 95% confidence intervals, and ICC's for the lifts and manual handling score are presented in table 1.

Insert table 1 about here

The ICC's for the three lifts show an excellent reliability (0.90 – 0.92), with the CI lower bounds being considered highly reliable. The limits of agreement of the three lifts ranged between $\pm 24\%$ of the mean score. This equates to ± 3 to 4kg variation in weight lifted. Figure 1 shows the distributions of the limits of agreement, where it can be seen that the majority of lifts were within ± 2.5 kg difference, however lower levels of agreement are seen with higher mean loads.

Insert figure 1 about here

A moderate reliability for the manual handling score (0.74) was identified. Further analysis, using a paired sample t-test, identified that manual handling scores in session two were rated significantly higher than the manual handling scores in session one. To more closely examine the components of the manual handling score (stance, posture, torque, leverage and pace), kappa and percentage agreements were calculated. The percentage agreement between the initial test and retest for these manual handling components was high in most cases, however the kappa results suggested more chance agreement. These results are presented in table 2.

To determine the internal consistency of the manual handling score the Cronbach's alpha was calculated for each lift indicating sufficient internal consistency, and that the items within the scale are measuring the same construct, in this instance manual handling technique. Together, the Cronbach's alpha for all manual handling was 0.92 indicating high internal consistency and therefore reliability of this rating scale.

Insert table 2 about here

DISCUSSION.

Literature has highlighted the importance of therapists using reliable and valid assessment tools to identify abilities and limitations of individuals [20]. Test-retest reliability is used to determine the consistency of a measure from one testing occasion to another[21]. Test-retest reliability for the lifting components of the WorkHab FCE was expressed by Intraclass Correlation Coefficients (ICC's) which is a measure of between- subject variance and within subject variance and is an accepted measure of reliability in relation to the

discriminative capacity of a test. To avoid bias, the test –retest time intervals must be far enough apart to avoid fatigue and close enough to avoid changes in performance. In this study a one week interval was used, with testing occurring at a similar time of day in both instances wherever possible to reduce bias.

For all three lifts evaluated in this study, the test-retest reliability was high (ICC's: 0.90-0.92). 95% limits of agreement were calculated as a descriptive measure of agreement with results being considered from a clinical rather than a statistical interpretation. Clinically a variation of ± 3 to 4kg in weight lifted maybe appropriate in some situations where heavier loads are lifted, however, when looking at the distribution of the limits of agreement (Figure 1), the majority of lifts were within ± 2.5 kg difference, which is clinically more acceptable. The results of any FCE need to be interpreted and applied to the specific tasks, job and workplace for the individual, using clinical judgement skills. Therefore, when considering the use of FCE to determine functional ability for return to work this variation can be regarded as acceptable and it implies that this is an acceptable clinical measure of agreement when looking at physical demands related to work tasks. This suggests that the administration procedures for the WorkHab FCE are dependable and the average performance by the subjects was relatively stable over the study period. Similar types of lifting tasks evaluated in other FCE's such as the Ergo-kit, Isernhagen and Physical Work Performance Evaluation, have reported substantial or acceptable test-retest reliability [2, 12, 19, 24, 30].

Whilst the manual handling scores indicated only a moderate level of reliability, there was a significant increase in scores from the first to the second assessment, which may be due to an improvement in the manual handling technique in the retest. The testing effect of practice with this sample suggested that the education provided and practiced in the initial test was learned and applied in the retest situation. Bloom suggests that timing is crucial in mastering learning, with weaker students needing more time to reach proficiency than more able students. In our study, a university population was used, therefore, these individuals, by nature of the environment sampled, are more able, suggesting the learning effect from the initial test was transferred to the retest resulting in an improvement in manual handling technique[31].

The percentage agreement results for the manual handling score components need to be considered in conjunction with the kappa results when determining test-retest reliability, taking into account any learning effects. The kappa scores were all in the poor to moderate range, although the percentage agreement was good. Both these calculations have limitations as percentage agreement does not take into account chance whereas the kappa does, and in the case of lack of variation in the spread of data across cells, large discrepancies may exist between the kappa score and the percentage of agreement, [20] as was the case in this study. This discrepancy can result from most agreement being limited to only one of the possible rating choices which means only one decision can make the difference between a poor and excellent reliability kappa score [32].

The other consideration in relation to the manual handling score is that from a clinical perspective, an improvement in this score is positive, as the individual has learned and is utilising appropriate lifting techniques. In practice, the administration of the WorkHab FCE requires feedback to the client to ensure appropriate lifting techniques are learned. Therefore, adequate test-retest reliability for these manual handling components can be assured from these results when applied to the practice context, given the overall results.

Results indicated that the WorkHab FCE manual handling scoring scale demonstrated sufficient internal consistency, where all the items within the scale for each lift measured the same construct of manual handling. This provides further evidence for clinicians about the reliability of the manual handling scoring scale.

One limitation of this study is that subjects were healthy individuals with no manual handling restrictions, which may have produced more positive reliability than if subjects were from an injured population. As the WorkHab FCE is typically conducted with injured workers, this needs to be taken into account, however, the WorkHab FCE is also used as a pre-employment functional capacity evaluation[19] and in this instance would be assessing healthy individuals, so the study results can be confidently applied with this population. This study used only one assessor who was a trained, experienced WorkHab FCE assessor, and data collected during the initial test was not reviewed prior to the retest occasion to reduce bias. The small

sample may also be considered a limitation, although sample size calculations indicated that this number provided us with adequate power to make generalizations across a healthy population. Further studies on an injured population would overcome the limitations of this study.

CONCLUSION.

The primary purpose of this research was to investigate the test –retest reliability for the floor to bench, bench to shoulder and bench to bench lift during the WorkHab FCE with healthy adults. The results of the lifting found substantial levels of test-retest reliability with this group. It can therefore be concluded, that these lifts in the WorkHab FCE are reliable in healthy adults. The manual handling scoring scale, as part of the WorkHab FCE was also investigated and had good internal consistency. However, the results relating to test-retest reliability showed moderate reliability as a result of an improvement in manual handling technique at the retest assessment. Further research is recommended to establish other forms of reliability and validity of this assessment tool, using a range of client samples. The results from this study contribute to the growing evidence of FCE's in practice and the importance of reliability and validity in work related assessments.

ACKNOWLEDGEMENT

The authors would like to thank the individuals who volunteered to participate in this study and WorkHab Australia who donated the equipment to complete the study. This study was made possible with the support of a University of Newcastle grant (No: GO187308).

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References:

- [1] King PM, Tuckwell N, Barrett TE. A critical review of functional capacity evaluations. *Physical Therapy*. 1998;78(8):852.
- [2] Gouttebauge V, Wind H, Kuijjer 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.
- [3] 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.
- [4] Gross D, Battie M, Asante A. Evaluation of a Short form Functional Capacity Evaluation: Less maybe Best. *Journal of Occupational Rehabilitation*. 2007;17:422-35.
- [5] Innes E, Straker L. Attributes of excellence in work related assessments. *Work*. 2003;20:63-76.
- [6] Jones T, Kumar S. Functional capacity evaluation of manual materials handlers: a review. *Disability and Rehabilitation*. 2003;25(4-5):179-91.
- [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] Reesink D, Jorritsma W, Reneman M. Basis for a Functional Capacity Evaluation Methodology for Patients with Work-related Neck Disorders. *Journal of Occupational Rehabilitation*. 2007;17(3):436-49.
- [15] Gouttebauge V, Wind H, Kuijjer 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.
- [16] Deen M, Gibson L, Strong J. A survey of occupational therapy in Australian work practice. *Work*. 2002;19:219-30.

- [17] Cotton A, Schonstein E, Adams R. Use of Functional Capacity Evaluations by rehabilitation providers in NSW. *Work*. 2006;26:287-95.
- [18] 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.
- [19] Legge J, Burgess-Limerick R. Reliability of the JobFit system Pre-employment Functional assessment tool. *Work*. 2007;28:299-312.
- [20] Innes E, Straker L. Reliability of work-related assessments. *Work*. 1999;13(2):107-24.
- [21] Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, New Jersey: Pearson Prentice Hall 2009.
- [22] Innes E, Straker L. Validity of work-related assessments. *Work*. 1999;13(2):125-52.
- [23] Soer R, Gerrits E, Reneman M. Test-retest reliability of a WRULD functional capacity evaluation in healthy adults. *Work*. 2006 2006///;26(3):273-80.
- [24] Tuckwell NL, Straker L, Barrett TE. Test-retest reliability on nine tasks of the Physical Work Performance Evaluation. *Work*. 2002 11;19(3):243-53.
- [25] Bradbury S, Roberts D. *Workhab Functional Capacity Evaluation Procedural Manual* WorkHab Australia 1998.
- [26] Holtgreffe K, Glenn T. Principles of aerobic exercise. In: Kisner C, Colby L, eds. *Therapeutic exercise: Foundations and techniques*. Philadelphia: F. A. Davis 2007:231-49.
- [27] Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*. 1999 06;8(2):135-60.
- [28] Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to practice*. 2nd ed. New Jersey: Prentice Hall Health 2000.
- [29] Depoy E, Gitlin L. *Introduction to Research. Understanding and applying Multiple Strategies*. 2nd ed. St Louis: Mosby 1998.
- [30] 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.
- [31] Bloom BS. *Human Characteristics and School Learning*. New York: McGraw-Hill 1976.
- [32] Brouwer S, Dijkstra P, Stewart R, Goeken L, Groothoff J, Geertzen J. Comparing self report, clinical examination and functional testing in the assessment of work-related limitations in patients with chronic low back pain. *Disability and Rehabilitation*. 2005;27(17):999-1005.

Table 1: Results of means, differences, standard deviations, 95% confidence intervals and ICC's for lifts.

Lift	Mean 1 ^a	SD 1	Mean 2 ^b	SD ^c 2	Mean dif	SD of mean dif	95% CI ^d of difference	Limits of agreement	Ratio of LoA & mean (%) ^e	ICC ^f	95% CI of ICC	Interpretation of ICC
Floor to Bench	17.0	3.9	16.8	3.6	0.18	2.1	0.68 to 1.04	-4.0 to 5.2	24%	0.92	0.82 to 0.96	Excellent
Bench to Shoulder	13.3	3.3	13.0	2.6	0.31	1.7	-0.41 to 1.03	-3.0 to 4.0	22.6%	0.90	0.78 to 0.96	Excellent
Bench to Bench	16.8	4.7	16.8	3.8	-0.02	2.5	-1.07 to 1.04	-3.7 to 2.2	21%	0.91	0.79 to 0.96	Excellent
Manual Handling score	15.7	2.04	16.7	1.3	-1.02*	1.3	-1.55 to -0.49	-	-	0.74	0.42 to 0.88	Moderate

^a: Mean 1 = group mean in first session (kg).

^b: Mean 2 = group mean in second session (kg).

^c: SD = Standard deviation.

^d: 95% CI = 95% Confidence interval.

^e: ratio between limits of agreement and mean score x100%

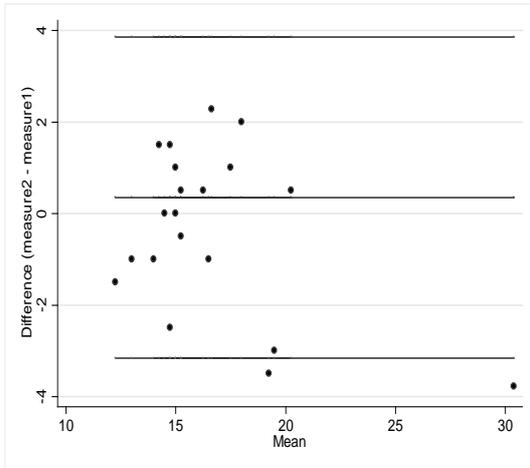
^f: ICC = Intraclass Correlation Coefficient (one way random).

*Significant (two tailed) at p<0.05

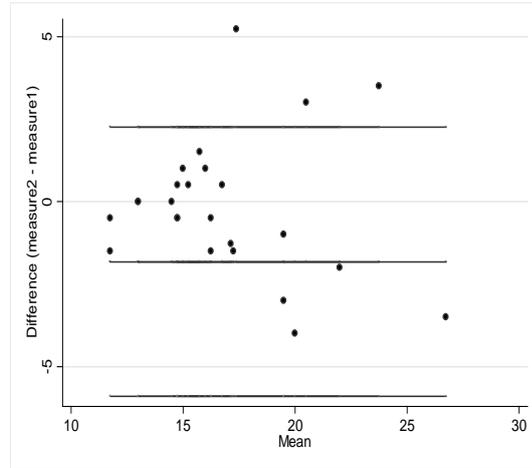
Table 2: Percentage agreement and Cronbach's alpha for the manual handling score components.

Lift	Manual Handling component	Percentage agreement between test & retest	Kappa	Cronbach's alpha.	Cronbach's alpha overall
Floor to Bench	Stance	87%	Moderate (0.47))0.86)
	Posture	92%	Moderate (0.52))
	Leverage	80%	Poor (0.32))
	Torque	84%	Poor (0.25))
	Pace	80%	Poor (0.31))
Bench to Shoulder	Stance	79%	Poor (0.30))0.85)0.92
	Posture	85%	Poor (0.24))
	Leverage	80%	Poor (0.30))
	Torque	81%	Poor (0.38))
	Pace	72%	Poor (-0.06))
Bench to Bench	Stance	85%	Moderate (0.51))0.82)
	Posture	85%	Moderate (0.41))
	Leverage	75%	Poor (0.33))
	Torque	84%	Poor (0.35))
	Pace	83%	Poor (0.17))

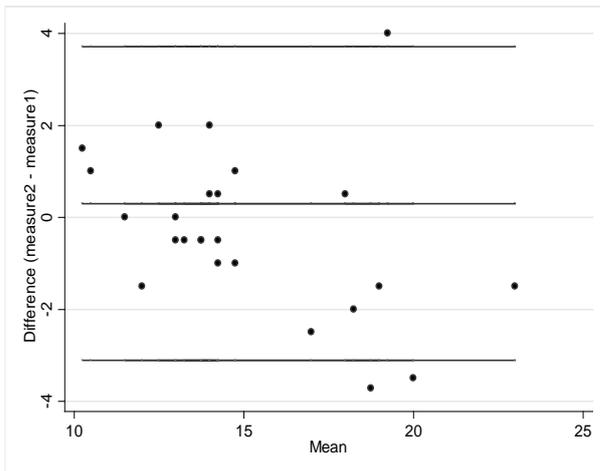
Figure 1: Limits of Agreement: Bench to bench, Floor to bench and Bench to shoulder lifts.



Bench to bench lift.



Floor to bench lift



Bench to shoulder lift