The reliability of determining effort level of lifting and carrying in a functional capacity evaluation

Reneman, M. F.; Jaegers, S. M.H.J.; Westmaas, M.; Göeken, L. N.H.

Published in:
Work

Publication date:
2002

Work 18 (2002) 23–27
IOS Press

The reliability of determining effort level of lifting and carrying in a functional capacity evaluation

M.F. Reneman[a,*], S.M.H.J. Jaegers[a], M. Westmaas[b] and L.N.H. Göeken[c]
[a]*University Rehabilitation Center Beatrixoord, Haren, The Netherlands*
[b]*Institute for Movement Sciences, University of Groningen, The Netherlands*
[c]*Department of Rehabilitation, University Hospital Groningen, Institute for Movement Sciences, University of Groningen, The Netherlands*

Received 15 March 2001
Accepted 14 May 2001

Abstract: *Objectives:* To establish inter- and intra-rater reliability of observations in a functional capacity evaluation.
Background: Functional capacity evaluations are used to assess a person's functional capacity as it relates to work. Lifting and carrying are important aspects of a functional capacity evaluation. An evaluator determines the patient's levels of effort through standardized observations. Questions remain with regards to the reliability of these observations.
Methods: Four healthy subjects were videotaped while performing two lifts and four carries with progressive loads. The videotape was scrambled randomly and viewed twice by 3 physical therapists and 2 occupational therapists. The evaluators determined the amount of effort it required (light, medium, heavy, and maximum). The inter- and intra-rater reliability of the observations was expressed by means of percentage agreement.
Results: Inter-rater reliability ranged 87–96%, intra-rater reliability ranged 93–97%.
Conclusion: The results indicate that by means of standardized observations, therapists can reliably determine effort level during lifting and carrying in healthy subjects, and thus affirm the findings of other studies of similar design.

Keywords: Interrater reliability, intrarater reliability, observational criteria, manual material handling, maximum effort

1. Introduction

Functional capacity evaluations (FCE's) are used to assess a person's functional capacity as it relates to work. FCE's are based on the Dictionary of Occupational Titles (DOT) [1,7,14], a publication of the United States Department of Labor. The DOT classifies work into five levels of physical demand: sedentary, light, medium, heavy and very heavy. It also identifies 20 job factors, two of which are lifting and carrying. A wide variety of devices and protocols are developed to measure a person's lifting and carrying capacities. The validity of these tests depends critically on the subject's effort during the evaluation [3,9]. The determination of whether a person has given maximal effort during the testing procedure appears to be difficult. The reliability of determining effort levels with the use of computerized lifting protocols has been questioned [3, 10]. Hazard et al. [3] compared several indices of subject effort, among which were isokinetic force/distance curve variations, peak force variations and heart rates. They conclude: "A trained observer is better to distinguish maximal from submaximal efforts than the most accurate physiologic index assessed in this study" and

*Address for correspondence: M.F. Reneman, MS, PT, University Rehabilitation Center Beatrixoord, P.O. Box 30002, 9750 RA Haren, The Netherlands. Tel.: +31 50 5338550; Fax: +31 50 5338550; E-mail: m.reneman@beatrixoord.nl.

"... the skilled evaluator remains a critical factor in validating lifting tests" [3].

Even though FCE's are being used routinely in the United States of America and also in many other countries worldwide [6,10], only three published studies were found with regards to the reliability of observations [5,8,13]. In these studies the observers reached substantial levels of inter- and intrarater agreement, as expressed in Cohen's Kappa scores, ranging from 0.62 to 0.88. Smith [13], however, studied the determination of safety in a floor-to-waist lift only. The inter-rater reliability study of Isernhagen et al. [5] was judged between raters and an expert observer. It is, however, not known what the qualifications of this expert observer are in order to be used as a "golden standard" in a scientific study. Whereas Smith and Isernhagen used video-observations, Lechner et al. [8] performed their study in a 'real-life' FCE situation. It is not unlikely that the determinations were based on information other than visual observations only.

The purpose of this study was to study the reliability of standardized observations. It duplicates in part the above-mentioned studies, however there are also differences. This study does not use a golden standard, all the lifts and carries are included, and the determinations are based on visual observations only. Additionally, observers were asked to determine four levels of effort, rather than two (maximal/submaximal).

2. Methods

2.1. Subjects

Two men and two women participated on a voluntary basis in this study. Their age ranged between 20 and 30 years. The subjects were healthy, had no current or previous complaints of back or neck pain and had normal cardiovascular resting values. All subjects provided informed consent for participation in this study.

2.2. Materials

Standardized materials used in the Isernhagen Work Systems FCE [6] were used in this study: a commercially available plastic receptacle (dimensions: depth $\times$ width $\times$ height $= 30 \times 40 \times 26$ cm) with handles on each side, a wall mounted system with in height adjustable shelves and metal weights of 4 and 2 kg. A toolbox like wooden receptacle (dimensions: $30 \times 30 \times 46$) was used for one-handed carries.

2.3. Procedures

After a general introduction of the procedures, signing the informed consent and measuring resting blood pressures and heart rate, the subjects were briefly verbally instructed on how to perform the lift or carry. The tester performed the lift or carry once to further explain the procedure. The subject then began to lift or carry the lightest load and progressed step by step to his endpoint. The loads were predetermined: men handled loads from 10 kg to 50 kg (5 increments of 10 kg), women handled loads from 6 to 30 kg (5 increments of 6 kg). The subjects were instructed to stop whenever they felt it became unsafe. Testing was ended either when the subject felt unsafe, or when the predetermined maximum weight was reached. For the purpose of this study the tester was not to interfere with the testing procedure on the basis of observations.

The subjects performed six material handling tasks: lifting low, lifting high, short carry, long carry two-handed, long carry right-handed and long carry left-handed.

- *Lifting low:* the receptacle was lifted from a 80 cm table, the subject turned 90 degrees towards the left, lowered the receptacle to the floor, briefly touching the floor, lifted toward an upright position, turned back 90 degrees and returned the receptacle to its original position. This was repeated 5 times within 90 seconds. Repetition 3, 4 and 5 are taped on video.
- *Lifting high:* the receptacle was lifted from a 80 cm table, the subject made one step backward, elevated the receptacle, rests it on the highest shelf positioned at a height so that the hands were at crown height, then returned the receptacle to its original position. This was repeated 5 times within 90 seconds. Repetition 3, 4 and 5 are taped on video.
- *Short carry:* the subject lifts the receptacle from the table (80 cm), turns 90 degrees, walks 1.2 meters (4 feet), turned 90 degrees, puts it on anther table (80 cm), then returns the receptacle to its original position. This is repeated 5 times within 90 seconds. Repetition 3, 4 and 5 are taped on video.
- *Long carry two-handed:* The receptacle was lifted from the table (80 cm), the subject turned approximately 180 degrees, carried the load over 16 meters and returned it to its original position within 90 seconds. Taped in full on video.

Table 1
Observational criteria to determine effort level [6, with permission]

Criteria	Maximal	Heavy	Moderate	Light
Muscle recruitment	Bulging of accessory muscles and trunk/neck stabilizers	Pronounced recruitment of accessory muscles and trunk/neck stabilizers	Recruitment of accessory muscles and trunk/neck stabilizers	Prime movers only; no accessory muscles, no trunk/neck stabilizers
Base of support	Very solid base	Wider base	Stable base	Natural stance
Posture	Marked counter balance	Increasing counter balance	Beginning of counter balance	Upright posture
Control and Movement pattern	Uses momentum in controled manner. Unable to control if weight is added.	Begins to use momentum. Difficult but not maximal.	Smooth movements	Easy movement patterns

– *Long carry one-handed* (*right and left*)*:* The "toolbox" was lifted with one hand from its position on the floor, carried over 16 meters, and returned within 90 seconds to its original position. Taped in full on video.

All procedures were copied into six clusters onto another videotape (low lifts in the first cluster, high lifts in the second, etc). Within each cluster, the magnitude of the load and the subjects were scrambled randomly. This means, for example, that the first frame could be subject 3 lifting a heavy weight, the second frame subject 1 lifting a light weight, the third frame subject 4 lifting maximally and so on.

2.4. Observers

Three physical therapists (PT's) and two occupational therapists (OT's) performed the ratings. Two PT's and one OT had completed a formal FCE training course. The other PT and OT were trained by one of the trained PT's. All therapists had actively participated in two 2-hour consensus meetings, which were held three months and just before the first observation. Four observers had performed between 1 and 5 FCE's, while one observer had performed approximately 100 FCE's. All therapists had at least 1 year of experience in occupational rehabilitation.

2.5. Observations

In theory, 120 different procedures could be taped (4 subjects $\times$ 6 procedures $\times$ 5 weight increments $=$ 120). The subjects, however, chose not to perform a total of 16 procedures due to their own judgement of heaving reached a safety endpoint. Thus, the tape consisted of 104 different procedures, which took 45 minutes to view. The first observation was performed in a single event where all five raters were present and simultaneously rated the video. The observers were blinded to each other's ratings. The second rating was performed individually and took place one week to two months after the first rating. The observers were asked to determine the effort level (light, medium, heavy, and maximum) using the observational criteria listed in Table 1.

2.6. Statistical analysis

Inter-rater reliability was calculated by comparing the amount of agreement between all paired observers (1-2, 1-3, 1-4, 1-5, 2-3, etc.). Intra-rater reliability was calculated similarly by comparing the scores of the first with the second observation within each observer (1-1, 2-2, etc.). Statistical analyses were performed using a computer program designed to compute agreement on nominal data [16,17].

3. Results

The results of the two rating sessions are presented in Table 2. The inter-rater reliability of the first rating session is presented as 'Inter I', the results of the second session as 'Inter II'. All but 1 of the determinations equaled or exceeded 90% agreement.

4. Discussion

This study was designed to investigate the inter- and intrarater reliability of determining effort level of healthy people during lifting and carrying by means of standardized visual observations only. When taken into account the predetermined standard of 90% agreement, all but one of the determinations exceed this level. The results of this study tend to affirm the findings of the studies of similar design mentioned in the introduction. Generalization of the results, however, should be made with great care due to the limitations of this study.

Table 2
The inter- and intrarater reliability of determining effort level expressed in percentage agreement (%)

	Inter session I	Inter session II	Intra
Lifting low	96	90	94
Lifting high	93	91	93
Carry short	95	92	96
Carry long	93	93	93
Carry left	95	93	97
Carry right	94	87	93

Levels of agreement when using nominal data are often expressed by means of a Cohen's Kappa coefficient. During statistical analysis of the results, however, it occurred that even though the percentage of agreement between or within raters were high (most in excess of 90%), the levels of agreement as expressed in a Cohen's Kappa score could be (extremely) low. For example, the inter-rater reliability of short carry appears to be as high as 93%, while expressed in Cohen's Kappa a score of -0.02 was computed. Further data analysis revealed the following explanation for this phenomenon. In a relatively small amount of total observations a difference in a single observation has high impact on the total score. Consistent with guidelines in literature [12,16] it was, therefore, decided not to use Cohen's Kappa as a measure for agreement in this study. A major advantage of the use of Cohen's Kappa is the ruling out of agreement by chance only (theoretically 25% in this study). Agreement expressed in percentages does not have the same statistical power as when expressed in a Cohen's Kappa and is known to overestimate the reliability [4]. The results of this study should be interpreted accordingly.

FCE's are often used to determine the functional capacity of people diagnosed with chronic non-specific pain to the locomotive system. The subjects used in this study were healthy young adults. The influence of pain behaviors on the observers' determinations is assumed to be non-existent. Generalization of the results of this study to a patient population should on these grounds also be performed with great care. Pain behaviors appear to be an important source of variance challenging reliability of FCE's, but are to our knowledge not assessed systematically in any of the well-known FCE's [9].

Appreciating the abovementioned limitations, the results of this study are consistent with reports of Smith [13], Isernhagen et al. [5] and Lechner et al. [8]. Even thought each study has its shortcomings, a considerable base of evidence is converging towards the point that the reliability of observations of this kind seems to be sufficient for its purpose, i.e. for clinical practice. There continues to be, however, gaps in knowledge that need to be filled in order to elevate the knowledge level regarding the reliability of observations during the testing of lifting and carrying capacity. Future studies need to incorporate more subjects in order to be able to rule out potential bias due to different lifting strategies. It should also use patients as subjects. There should be a considerable variance in effort levels in order to be able to use stronger statistical measures. The validity of the observations should also be addressed. This and other studies may have demonstrated the ability of therapists to reliably observe effort levels, it has yet to be determined whether the used (operational) definitions of maximal effort truly represent maximum effort in healthy subjects and in patients. Next to the observations of behaviors concurring with (increased) physical effort, the possibility of systematically assessing pain behaviors during functional capacity evaluations needs to be explored.

References

[1] E. Abdel-Moty, D.A. Fishbain, T.M. Kahlil, S, Sadek, R. Cutler, R. Steele Rosomoff AND H.L. Rosomoff, Functional Capacity and Residual Functional Capacity and their utility in Measuring Work Capacity, *Clin. J. Pain* **9** (1993), 168–173.
[2] W.J.J. Assendelft, B.W. Koes, P.G. Knipschild and L.M. Bouter, The relationship between methodological quality and conclusions in reviews of spinal manipulation, *JAMA* **274** (1995), 1942–1948.
[3] G.R. Hazard, V. Reeves and J.W. Fenwick, Lifting capacity; indices of subject effort, *Spine* **17** (1992), 1065–1070.
[4] E. Innes and L. Straker, Reliability of work-related assessments, *Work* **13** (1999), 107–124.
[5] S.J. Isernhagen, D.L. Hart and L.M. Matheson, Reliability of independent observer judgements of level of lift effort in a kinesiophysical Functional Capacity Evaluation, *Work* **12** (1999), 145–150.
[6] S.J. Isernhagen, *Isernhagen Work Systems; FCE Training handbook*, Not dated. Copyrighted material. Published with permission.
[7] P.M. King, N. Tuckwell and T.E. Barett, A critical review of Functional Capacity Evaluations, *Physical Therapy* **8** (1998), 852–866.
[8] D.E. Lechner, R.J. Jackson, D.L. Roth and K.V. Straaton, Reliability and validity of a newly developed test of physical work performance, *JOM* **36** (1994), 997–1004.
[9] D.E. Lechner, S.F. Bradbury and L.A. Bradley, Detecting sincerity of effort: a summary of methods and approaches, *Physical Therapy* **8** (1998), 867–888.
[10] M. Newton and G. Waddell, Trunk strength testing with isomachines; Part 1: Review of a decade of scientific evidence, *Spine* **8** (1993), 801–811.
[11] M.F. Reneman, S.M.H.J. Jaegers, C. Muskee, H.T.H. Schroer and L.N.H. Goeken, Functional Capacity Evaluation: toepassing in Nederland? *TBV* **5** (1997), 139–146.

[12] H.J.A. Schouten, Nominal scale agreement between observers, *Psychometrika* **5** (1986), 453–466.
[13] R.L. Smith, Therapists' ability to identify safe maximum lifting in low back patients during functional capacity evaluations, *J Orthop Sports Phys Ther* **19** (1994), 277–281.
[14] A.K. Tramposh, The functional capacity evaluation: Measuring maximal work abilities, *Occupational medicine: State of the Art Reviews* **7** (1993), 113–124.
[15] R. Popping, *Computing agreement on nominal data. The computer program Agree 6.0*, ProGamma, Groningen, the Netherlands, 1995.
[16] R. Popping, Agree 6 for nominal scale agreement, *ProGamma Info*, Spring, 1996, pp. 12–13.
[17] R. Popping, Personal communication, June 2001.